

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s): NAKAMURA, et al.
Serial No.: Not yet assigned
Filed: August 27, 2003
Title: STORAGE SYSTEM
Group: Not yet assigned

LETTER CLAIMING RIGHT OF PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

August 27, 2003


Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, the applicant(s) hereby claim(s) the right of priority based on Japanese Patent Application No.(s) 2003-199581, filed July 22, 2003.

A certified copy of said Japanese Application is attached.

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP



James N. Dresser
Registration No. 22,973

JND/alb
Attachment
(703) 312-6600

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日 2003年 7月22日
Date of Application:

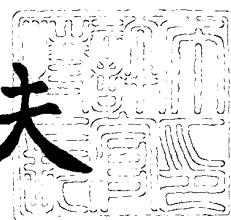
出願番号 特願2003-199581
Application Number:
[ST. 10/C]: [JP2003-199581]

出願人 株式会社日立製作所
Applicant(s):

2003年 8月14日

特許庁長官
Commissioner,
Japan Patent Office

今井康夫



出証番号 出証特2003-3065704

【書類名】 特許願

【整理番号】 K03008741A

【あて先】 特許庁長官殿

【国際特許分類】 G06F 12/16

【発明者】

 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所システム開発研究所内

 【氏名】 中村 崇仁

【発明者】

 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所システム開発研究所内

 【氏名】 島田 健太郎

【特許出願人】

 【識別番号】 000005108

 【氏名又は名称】 株式会社 日立製作所

【代理人】

 【識別番号】 100075096

 【弁理士】

 【氏名又は名称】 作田 康夫

【手数料の表示】

 【予納台帳番号】 013088

 【納付金額】 21,000円

【提出物件の目録】

 【物件名】 明細書 1

 【物件名】 図面 1

 【物件名】 要約書 1

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 記憶装置システム

【特許請求の範囲】

【請求項 1】

計算機と接続される記憶装置システムであって、

第一の制御部、第二の制御部、第三の制御部及び複数の記憶装置を有し、

前記第一の制御部、前記第二の制御部及び前記第三の制御部の各々はメモリを有し、

前記第一の制御部は、前記計算機から受信したデータを当該第一の制御部が有するメモリと前記第二の制御部が有するメモリに格納することを特徴とする記憶装置システム。

【請求項 2】

前記第二の制御部が使用不能となった際には、前記第一の制御部は、前記計算機から受信したデータを、当該第一の制御部が有するメモリと前記第三の制御部が有するメモリに格納することを特徴とする記憶装置システム。

【請求項 3】

前記第二の制御部は、前記計算機から受信したデータを当該第二の制御部が有するメモリ及び前記第一の制御部が有するメモリに格納することを特徴とする請求項 2 記載の記憶装置システム。

【請求項 4】

前記第二の制御部が使用不能となった際には、前記第一の制御部が前記第二の制御部に代わって前記計算機からデータを受け付け、当該第一の制御部は、前記受け付けたデータを当該第一の制御部が有するメモリ及び前記第三の制御部が有するメモリに格納することを特徴とする請求項 3 記載の記憶装置システム。

【請求項 5】

更に第四の制御部を有し、前記第四の制御部はメモリを有し、

前記第二の制御部が使用不能となった際には、前記第一の制御部が前記第二の制御部に代わって前記計算機からデータを受け付け、当該第一の制御部は、前記受け付けたデータを当該第一の制御部が有するメモリ及び前記第四の制御部が有

するメモリに格納することを特徴とする請求項 3 記載の記憶装置システム。

【請求項 6】

前記第一の制御部と前記第二の制御部は別電源から電源供給を受けていることを特徴とする請求項 5 記載の記憶装置システム。

【請求項 7】

前記第三の制御部と前記第四の制御部は別電源から電源供給を受け、かつ、前記第一の制御部と前記第三の制御部は別電源から電源供給を受けていることを特徴とする請求項 6 記載の記憶装置システム。

【請求項 8】

前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部の各々を接続するスイッチを有し、前記各制御部は、前記スイッチを介して前記計算機と接続されることを特徴とする請求項 5 記載の記憶装置システム。

【請求項 9】

インターフェース部を有し、
前記スイッチは前記インターフェースを介して前記計算機と接続され、
前記インターフェース部、前記スイッチ、前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部で一つの制御装置となっていることを特徴とする請求項 8 記載の記憶装置システム。

【請求項 10】

管理用装置を有し、
前記管理用装置は、前記スイッチを介して前記インターフェース部、前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部に接続されることを特徴とする請求項 9 記載の記憶装置システム。

【請求項 11】

前記管理用装置は、当該記憶装置システムにおける前記記憶装置と前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部との関係を示す情報を有し、

前記インターフェース部、前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部は、前記情報に基づいて前記データの格納を実行す

ることを特徴とする請求項 10 記載の記憶装置システム。

【請求項 12】

前記情報には、前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部のいずれかに障害が発生した際に、前記障害が発生した制御部の代替となる制御部及び前記障害が発生した制御部が受取るデータの複製を格納するメモリを有する制御部を指定する情報が含まれ居ていることを特徴とする請求項 11 記載の記憶装置システム。

【請求項 13】

前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部のいずれかに障害が発生した場合、前記管理用装置が前記障害を検出して前記情報を前記障害の状態に応じて変更し、前記インターフェース部、前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部に前記障害の発生及び前記情報の変更を通知し、前記インターフェース部、前記第一の制御部、第二の制御部、第三の制御部及び前記第四の制御部は前記変更された情報に基づいて動作することを特徴とする請求項 12 記載の記憶装置システム。

【請求項 14】

更に前記障害が回復した場合、前記管理用装置は、前記障害の回復を前記インターフェース部、前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部に通知することを特徴とする請求項 13 記載の記憶装置システム。

【請求項 15】

前記インターフェース部、前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部の各々は、当該記憶装置システムにおける前記記憶装置と前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部との関係を示す情報を有し、

前記インターフェース部、前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部は、前記情報に基づいて前記データの格納を実行することを特徴とする請求項 9 記載の記憶装置システム。

【請求項 16】

前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部のいずれかに障害が発生した場合、前記インターフェース部が前記障害を検出し前記情報を前記障害の状態に応じて変更し、前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部に前記障害の発生及び前記情報の変更を通知し、

前記第一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部は、前記変更された情報に基づいて動作することを特徴とする請求項 15 記載の記憶装置システム。

【請求項 17】

前記第二の制御部に障害が発生した際、前記第一の制御部は、その時点で該第一の制御部に格納されているデータを前記第三の制御部のメモリに転送することを特徴とする請求項 4 記載の記憶装置システム。

【請求項 18】

前記第二の制御部に障害が発生した際、前記第一の制御部は、その時点で該第一の制御部に格納されているデータを前記記憶装置に格納することを特徴とする請求項 4 記載の記憶装置システム。

【請求項 19】

前記第二の制御部に障害が発生した際、前記第三の制御部及び前記第四の制御部は、前記第一の制御部からのデータを受け付けるために、各々のメモリに格納されているデータの一部を前記記憶装置に格納することを特徴とする請求項 4 記載の記憶装置システム。

【請求項 20】

前記第一の制御部及び前記第二の制御部に障害が発生した場合、前記インターフェース部は、前記第一の制御部が管理する前記記憶装置への前記計算機からのアクセスに対して、エラー報告をすることを特徴とする請求項 9 記載の記憶装置システム。

【請求項 21】

管理用装置を有し、

前記管理用装置は、前記スイッチを介さずに前記インターフェース部、前記第

一の制御部、前記第二の制御部、前記第三の制御部及び前記第四の制御部に接続されることを特徴とする請求項 9 記載の記憶装置システム。

【発明の詳細な説明】

【0 0 0 1】

【発明の属する技術分野】

本発明は、キャッシュメモリを備える記憶装置システムとそのキャッシュメモリの制御方法に関する。

【0 0 0 2】

【従来の技術】

複数の記憶装置を有する記憶装置システム（以下「ストレージシステム」）の性能を向上するための技術として、揮発性の半導体記憶部（以下「キャッシュメモリ」）をストレージシステムに導入する技術が知られている。

【0 0 0 3】

キャッシュメモリを有するストレージシステムは、データの書き込み要求に対して、キャッシュメモリにデータを書き込んだ時点でデータ書き込みを要求した計算機（以下「コンピュータ」又は「ホストコンピュータ」と称する）に書き込み完了の応答を返し、それと非同期に記憶装置にデータを書き込む。キャッシュメモリへのデータの書き込み速度は記憶装置（ここではディスク装置等）よりも高速なため、ストレージシステムは、より高速にホストコンピュータに応答を返すことができる。

【0 0 0 4】

しかし、データが記憶装置に書き込まれるまでの間は、キャッシュメモリにしか最新のデータが存在しないので、ストレージシステムでは、キャッシュメモリの信頼性を向上させる必要がある。

【0 0 0 5】

キャッシュメモリの信頼性を向上するための技術として、キャッシュメモリを冗長化する方法が知られている。冗長化の方法としては、例えば複数のキャッシュメモリにデータのコピーを格納すること（ミラーリング）や、特許文献 1 に開示された R A I D 構成のキャッシュメモリがある。

【0006】

更に、キャッシュメモリの障害等でキャッシュメモリの冗長度が失われた場合においてもストレージシステムの信頼性を維持するため、各書き込み要求に対して必ず記憶装置にデータを保存する制御方法（「ライトスルー制御」）が知られている。しかし、ライトスルー制御によって信頼性は維持されるが、上述したキャッシュメモリの利点は失われ、キャッシュメモリを有していても、書き込み要求に対する応答速度はキャッシュメモリを有しない場合と同等となってしまう。

【0007】

そこで、ライトスルー制御を必要としないように、キャッシュメモリの冗長度を増加させる技術が考案されている。例えば、キャッシュメモリの予備を備えることや、特許文献2に開示されているような、3個以上のキャッシュメモリを備え、障害が発生したキャッシュメモリが担当していた領域に書き出すライトデータを残りのキャッシュメモリで分担する等の技術である。

【0008】**【特許文献1】**

特開平9-265435号公報

【特許文献2】

特開2001-344154号公報

【0009】**【発明が解決しようとする課題】**

現在、このようなストレージシステムを更に大規模に構成する要求が高まっている。しかし、従来技術では、キャッシュメモリを一元的に使用する。このため、ストレージシステムの構成規模が大きくなるにつれてキャッシュメモリやキャッシュメモリを管理するための情報にアクセスが集中し、単にキャッシュメモリを有するだけではストレージシステムのスループット性能の維持が困難になるという問題がある。

【0010】

また、キャッシュメモリに障害が発生した場合におけるストレージシステムの信頼性とライト性能維持に関しても上述の問題と同様の問題がある。即ち、前記

特許文献 2 記載の技術等はキャッシュメモリを一元的に使用しているので、構成規模が大きくなるにつれ、キャッシュメモリの障害時にキャッシュメモリやキャッシュメモリを管理するための情報にアクセスが集中し、単にキャッシュメモリを有するだけではスループット性能の維持が困難になり、構成規模の拡大と障害時の信頼性等の両立は困難である。

【0011】

即ち、本発明の目的は、キャッシュ障害発生時においてもライトアクセス応答速度と信頼性を維持する、大規模構成可能なストレージシステムおよびその制御方法を提供することである。

【0012】

【課題を解決するための手段】

上記目的を達成するために、本発明は以下の構成を有する。即ち、複数の制御部及び記憶装置を有する記憶装置システムである。更に、複数の制御部は各々がメモリ、例えばキャッシュメモリを有する。そして、このような構成の記憶装置システムにおいて、複数の制御部のうち第一の制御部は、記憶装置システムと接続される計算機からデータを受信した際には、第一の制御部が有するメモリ及び他の制御部（以下「第二の制御部」）が有するメモリに受信したデータを格納し、その後、記憶装置へデータを転送する。

【0013】

又、上記構成において第二の制御部に障害が発生した場合には、第一の制御部は、新たに第三の制御部のメモリに、計算機から受信したデータの複製を格納する。

更に、第二の制御部は、計算機から受信したデータを第一の制御部が有するメモリ及び第二の制御部が有するメモリに格納する構成としても良い。

【0014】

更に、第二の制御部に障害が発生した場合には、ペアとして指定されている第一の制御部が第二の制御部の処理を代行する構成としても良い。この場合、第一の制御部は、第二の処理部の代行中に計算機から受取ったデータの複製を、ペアではない他の制御部が有するメモリに格納する。

【0015】

又、ペアとなる第一の制御部と第二の制御部とは各々別電源から電源供給を受ける構成としても良い。

更に、複数の制御部は、スイッチを介して相互に接続される構成としても良い。更に、各制御部は、計算機とインターフェース部を介して接続する構成とすることもできる。

又、記憶装置システムは管理用装置を有し、この管理用装置が複数の制御部と記憶装置との対応関係を示す情報を有し、各制御部は、この情報に基づいて動作を行う構成とすることもできる。又、記憶装置システムが管理用装置を有さず、上記の情報を各制御部が有する構成とすることもできる。

【0016】

更に、ペアとなる制御部には、同じ記憶装置が接続される構成とする。

又、制御部の障害の発見を管理用装置が行う構成でも、他の制御部又はインターフェース部が、制御部の障害を検出する構成とすることもできる。

【0017】**【発明の実施の形態】**

以下、本発明の実施の形態を、図面を参照して説明する。ただし、本発明が以下に開示される実施形態に限定されないのは言うまでもない。

図1は、本発明を適用したストレージシステムの第1の実施形態を示した図である。ストレージシステムは、ディスク制御装置5及び複数のディスク装置4を有する。尚、ディスク装置4とは、ハードディスクドライブやCD、DVD等の不揮発性の記憶媒体を有する記憶装置である。ディスク制御装置5は通信線（以下「チャンネル」）61を介してホストコンピュータ6に接続されている。又、ディスク制御装置5とディスク装置4とは通信線（以下「ディスク側チャンネル」）41を介して相互に接続されている。ホストコンピュータ6は、チャンネル61、ディスク制御装置5及びディスク側チャンネル41を介して、ディスク装置4との間でデータを送受信する。

【0018】

チャンネル61及びディスク側チャンネル41では、例えばSCSI(Small

Computer System Interface)やファイバチャネルなどのプロトコルが用いられる。また、チャネル61は、ファイバチャネルケーブルと複数のファイバチャネルスイッチなどで構成されるSAN (Storage Area Network) で構成されていても良い。

【0019】

ディスク装置4は2つのポートを有し、その各々は、別々のディスク側チャネル41を介してディスク制御装置5に接続されている。これにより、ディスク制御装置5は、同一のディスク装置4に複数の経路（以下「パス」）を介してアクセスすることが可能である。

【0020】

ディスク制御装置5は、複数の電源部A511、B512、複数のホストアダプタ1、複数のキャッシュアダプタ3及び管理アダプタ7を有する。さらに複数のホストアダプタ1、複数のキャッシュアダプタ3及び管理アダプタ7は、内部スイッチ2を介して相互に接続されている。内部スイッチ2とキャッシュアダプタ3等との接続には、通信線である内部結合21が用いられる。本実施形態では内部スイッチ2と各要素との間の内部結合21は一本であるが、障害の発生に対する冗長性の確保、データ送信の通信帯域の確保、もしくは通信で使用する異なったパケット長に対応するため、内部スイッチ2と各要素との間の内部結合21が複数備えられてもよい。

【0021】

尚、管理アダプタ7は、ホストアダプタ1やキャッシュアダプタ3と、内部結合21とは異なるネットワークで接続されていても良い。これにより、データ転送に関わるネットワークと、システム管理用の情報の授受に関するネットワークを分離することができる。

【0022】

ホストアダプタ1は、チャネル1を介してホストコンピュータ6からのアクセス要求を受領し、所持する管理テーブル11に基づいてアクセス要求の解析を行い、内部結合21を介して適切なキャッシュアダプタ3と通信し、ホストコンピュータ6に応答を返すインターフェース装置である。

【0023】

キャッシュアダプタ 3 は、ディスク側チャネル 41 を介してディスク装置 4 と接続され、内部スイッチ 2 を介してホストアダプタ 1 や他のキャッシュアダプタ 3 と通信する。又、キャッシュアダプタ 3 は、ホストアダプタ 1 からの通信に基づいてディスク装置 4 からのデータの読み出し又はディスク装置 4 へのデータの書き込みを制御する。また、キャッシュアダプタ 3 は、自身が有するキャッシュメモリ 32 を制御し、キャッシュメモリ 32 へのデータの格納等を行う制御装置である。

【0024】

尚、キャッシュアダプタ 3 は、基本的には、自己に接続されているディスク装置 4 に格納されるデータの読み出し又は書き込みに関わるデータのみしかキャッシュメモリ 32 に格納しない。言い換えると、他のキャッシュアダプタ 3 に管理されているディスク装置 4 のデータは、通常に使用されるデータとしては、他のキャッシュアダプタ 3 のキャッシュメモリ 32 には格納されない。

【0025】

キャッシュアダプタ 3 は、ディスク装置 4 に対する冗長化の制御（例えば、各 RAID レベルの冗長化）も行う。更に、キャッシュアダプタ自身を冗長化するため、あるキャッシュアダプタ 3 と一つのポートで接続される各ディスク装置 4 のもう一方のポートは、そのキャッシュアダプタ 3 とペアになる別のキャッシュアダプタ 3 に接続される。

【0026】

尚、本実施形態では、複数のキャッシュアダプタ 3 のペアが一つのディスク制御装置 5 内に格納されている構成について説明するが、他の構成として、一つのペアとそのペアに共有されるディスク装置 4 とで一つの装置を構成し、これらの装置がスイッチ 20 を介して相互に接続される構成でも良い。この場合、管理用の装置（管理アダプタ）がスイッチ 20 を介して各ペアを管理する。

【0027】

管理アダプタ 7 は、ストレージシステムの構成についての情報が登録されたマスタ管理テーブル 71 を備える。管理アダプタ 7 は、ストレージシステムの構成

等が変更された場合などにマスタ管理テーブル 71 の内容を変更し、必要な情報をホストアダプタ 1 やキャッシュアダプタ 3 に配信する。

【0028】

電源部 A 511 及び電源部 B 512 は、それぞれ商用電源など外部電源（図示せず）に接続され、ストレージシステムに電力を供給する。電源事故に備え、電源部 A 511 及び電源部 B 512 は各々別系統の外部電源に接続されることが望ましい。又、本実施形態では、キャッシュアダプタ 3 の冗長性を確保するため、あるキャッシュアダプタ 3 とペアとなるキャッシュアダプタ 3 は、各々別の電源部 A 511 及び電源部 B 512 から電力を供給される。

【0029】

尚、他の装置構成として、ホストアダプタ 1 が存在せず、ホストアダプタ 1 が所持していた管理テーブル 11 を備えるスイッチ 20 により各キャッシュアダプタ 3 が相互に接続される構成もある。この場合、スイッチ 20 は複数のチャンネル 61 と接続される。また、スイッチ 20 は、チャンネル 61 毎に管理テーブル 11 を備え、各々の管理テーブル 11 に基づいて、ホストコンピュータ 6 のアクセス要求をキャッシュアダプタ 3 に転送する。

【0030】

また、スイッチ 20 が使用される場合は、ホストアダプタ 1 が行っていたホストコンピュータ 6 との間の通信やプロトコル変換等は、キャッシュアダプタ 3 が行う。

【0031】

図 2 は、キャッシュアダプタ 3 の構成例を表した図である。キャッシュアダプタ 3 は、キャッシュメモリ 32、内部結合 21 と接続される内部結合インタフェース（以下「I/F」）部 33、ディスク側チャンネル 41 と接続されるディスク側チャンネル I/F 部 34、プロセッサ 37、制御メモリ 36 及びプロセッサ周辺制御部 35 を有する。

【0032】

キャッシュメモリ 32、内部結合 I/F 部 33 及びディスク側チャンネル I/F 部 34 はキャッシュデータバス 38 により相互に接続されている。内部結合 I/F 部 33

及びディスク側チャンネルI/F部34は、装置間の直接データ転送（DMA）を行うことができる。具体的には、内部結合I/F部33は、内部結合21を介してホストコンピュータ6から受領したデータを、キャッシュデータバス38を介してキャッシュメモリ32に格納する。ホストコンピュータ6からリード要求を受領したら、内部結合I/F部33は、キャッシュメモリ32に格納されているデータをキャッシュデータバス38を介して取り出し、内部結合21を通じてホストアダプタ1に転送する。

【0033】

ディスク側チャンネルI/F部34は、キャッシュメモリ32に格納されたデータをキャッシュデータバス38を介して取り出し、ディスク側チャンネル41を介してディスク装置4に格納する（以下「デステージング」）。また、ディスク側チャンネルI/F部34は、ディスク側チャンネル41を介してディスク装置4に格納されたデータを取り出し、キャッシュデータバス38を介してキャッシュメモリ32に格納する（以下「ステージング」）。

【0034】

内部結合I/F部33及びディスク側チャンネルI/F部34は、制御データバス39を介したプロセッサ37の制御に基づいて、上述のステージング及びデステージング等の処理を実行する。

【0035】

プロセッサ37は、メモリ制御回路やバス制御回路を含むプロセッサ周辺制御部35を介して制御メモリ36及び制御データバス39に接続される。制御メモリ36には、管理テーブル31、制御プログラム361及びディレクトリ情報362が格納されている。

【0036】

管理テーブル31には、ホストアダプタ1から指定される論理デバイス（以下「LDEV」）、複数のディスク装置4を仮想的に1つのデバイスとして管理する場合の仮想デバイス（以下「VDEV」）及びLDEVに格納されるデータを冗長化（ここでは複製）して格納するキャッシュアダプタ3（以下「バックアップキャッシュアダプタ」）との対応関係を示す情報が登録されている。

【 0 0 3 7 】

制御プログラム 3 6 1 は、プロセッサ 3 7 がキャッシュアダプタ 3 の有する各構成要素の制御を実行する際にプロセッサ 3 7 で実行されるプログラムである。ディレクトリ情報 3 6 2 は、アクセス対象となるデータのキャッシュメモリ 3 2 での有無やキャッシュメモリ 3 2 でのアドレス等、データのキャッシュメモリ 3 2 への格納状況を示す情報である。

【 0 0 3 8 】

図 3 は、ホストアダプタ 1 が保持する管理テーブル 1 1 及びキャッシュアダプタ 3 が保持する管理テーブル 3 1 の内容例を示した図である。管理テーブル 1 1 は、複数のエントリを有し、各エントリはフィールド 1 1 1 及び 1 1 2 を有する。フィールド 1 1 1 には、ホストコンピュータ 6 がアクセスの際に指定する論理ユニット番号（L U 番号）が登録される。

【 0 0 3 9 】

フィールド 1 1 2 は、キャッシュアダプタ 3 に関する情報が格納されるフィールド 1 1 2 1、1 1 2 2 及び 1 1 2 3 のサブフィールドを有する。サブフィールド 1 1 2 1 には、フィールド 1 1 1 に登録された L U に対応する、キャッシュアダプタ 3 が管理する論理デバイス番号（L D E V 番号）の値が登録される。サブフィールド 1 1 2 2 には、ステージング、デステージングを行うキャッシュアダプタ 3、即ちディスク装置 4 に対して通常のデータの書き込み、読み出しを行うキャッシュアダプタ（以下「マスタキャッシュアダプタ」）を示す情報が登録される。サブフィールド 1 1 2 3 には、サブフィールド 1 1 2 2 に登録されたマスタキャッシュアダプタのキャッシュメモリ 3 2 に格納されたデータの冗長化を行うバックアップキャッシュアダプタを示す情報が登録される。

【 0 0 4 0 】

管理テーブル 3 1 には、上述したように、L D E V、V D E V 及びバックアップキャッシュアダプタとの対応関係を示すマッピング情報が格納される。管理テーブル 3 1 も複数のエントリを有し、各エントリは、フィールド 3 1 1、3 1 2、3 1 3 及び 3 1 4 を有する。フィールド 3 1 1 には、一つのエントリに対応する L D E V の L D E V 番号を示す情報が登録される。フィールド 3 1 2 には、フ

フィールド 311 に登録された LDEV のデータを冗長化するバックアップキャッシュアダプタを示す情報が登録される。

【0041】

フィールド 313 には、フィールド 311 に登録された LDEV に対応する VDEV 番号を示す情報が登録される。フィールド 314 には、フィールド 311 に登録された LDEV が、対応する VDEV のどの部分に割り当てられているかを示す仮想デバイスアドレス（以下「VDEV アドレス」）を示す情報が登録される。尚、VDEV は、ストレージシステムの管理者が、SVP（図示せず）や管理アダプタ 7 に接続したコンソールを通じて、またはチャンネルの特殊なコマンドを送付することで指定する。

【0042】

なお、フィールド 312 に登録されるバックアップキャッシュアダプタがその管理テーブル 31 を保持するキャッシュアダプタ 3 であれば、そのキャッシュアダプタ 3 は、対応するフィールド 311 に登録された LDEV 番号で指定される LDEV に関してバックアップキャッシュアダプタとしてライトデータ冗長化の処理を行う。具体的には、バックアップキャッシュアダプタは、ホストアダプタ 1 やマスタキャッシュアダプタ 3 より冗長化するライトデータを受領し、そのデータをキャッシュメモリ 32 に保存する。

【0043】

図 4 は、記憶装置システムがホストコンピュータ 6 からリード要求を受信した場合の各アダプタでの処理手順を示すフローチャートである。

まず、ホストアダプタ 1 が、チャンネル 61 を介してホストコンピュータ 6 よりリード要求を受領する。以下、ホストアダプタを HA、マスタキャッシュアダプタを CA (m) と記述する（ステップ 2001）。

【0044】

リード要求を受信した HA 1 は、管理テーブル 11 より、リード要求で指定される LU 番号に対応する LDEV 番号及び CA (m) の情報を検索する（ステップ 2002）。その後、HA 1 は、内部結合 21 を介して、検索された CA (m) に対し、内部リード要求を送信する。ここで「内部リード（ライト）要求」と

は、ホストアダプタ 1 とキャッシュアダプタ 3 との間でやり取りされるデータ読み出し（データ書き込み）のメッセージである（ステップ 2003）。

【0045】

内部リード要求を受領した CA (m) は、内部リード要求に含まれているアドレス、サイズなどに基づいて、ディレクトリ情報 362 よりリード要求に対応するデータがキャッシュメモリ 32 中に存在するか判定（以下「キャッシュヒット判定」）する（ステップ 2004）。判定の結果、キャッシュメモリ 32 中に該当するデータが存在しない（以下「キャッシュミス」）場合、CA (m) は、ディスク装置 4 より該当するデータをステージングしてキャッシュメモリ 32 に格納し、該当するディレクトリ情報 362 を更新する（ステップ 2005）。

【0046】

ステップ 2005 の処理後又はステップ 2004 でキャッシュメモリ 32 に該当するデータが存在すると判断した場合、CA (m) は、該当するデータをキャッシュメモリ 32 より内部結合インタフェース部 33 を介して読み出し、内部リード要求を送信してきた HA 1 に転送する（ステップ 2006）。

データを受信した HA 1 は、ホストコンピュータ 6 に受信したデータを応答する（ステップ 2007）。

【0047】

図 5 は、ストレージシステムがホストコンピュータ 6 からデータのライト要求を受信した際の処理の流れを示すフローチャートである。以下、バックアップキャッシュアダプタを CA (b) と記述している。

【0048】

チャネル 61 を通じてホストコンピュータ 6 よりライト要求を受領した HA 1 は、管理テーブル 11 よりライト要求に含まれる LU 番号に対応する LDEV 番号、CA (m) 及び CA (b) の情報を検索する。

【0049】

その後、HA 1 は、内部結合 21 を介して、検索された CA (m) に対し内部ライト要求を送信する（ステップ 2101）。内部ライト要求を受信した CA (m) は、キャッシュメモリ 32 にライト要求に対応するデータ（以下「ライトデ

ータ」) を格納できる領域があるかディレクトリ情報 362 より判定する (ステップ 2104)。

【0050】

格納できる領域が無い場合、CA (m) は、LRU アルゴリズム等に基づき、どの LDEV および LDEV アドレスに該当するキャッシュメモリ 32 中のデータをディスク装置 4 に書き込むか決定し、そのデータをディスク装置 4 に書き込んだ後、該当領域を無効としてライトデータを格納できる領域を確保する。さらに CA (m) は、CA (b) に、無効としたデータの LDEV 番号および LDEV アドレスを通知する。

【0051】

通知を受けた CA (b) は、該当データを無効とし、ライトデータを格納できる領域を確保する。その後、CA (b) は、CA (m) に対して領域確保の通知を行う。なお、CA (b) のキャッシュメモリ 32 には、ライトデータに関しては CA (m) と同じデータ (アドレスは異なってもよい) が格納されているので、CA (b) における格納領域の判定には、ステップ 2104 で行われた判定結果がそのまま適用できる (ステップ 2105)。

【0052】

ステップ 2105 の後、又はステップ 2104 で格納領域があると判断された場合には、CA (m) は、CA (b) に、ステップ 2105 で受領した内部ライト要求に対応した内部メッセージである内部バックアップライト要求を送信する。CA (m) および CA (b) の各々は、ライトデータの受け入れが可能な状態になったら、内部結合 21 を介して、内部ライト要求を送信した HA1 に対し内部メッセージである内部ライト準備応答を送信する。尚、CA (b) は、HA1 に内部ライト準備応答を送信する代わりに、CA (m) に内部ライト準備応答を送信し、CA (m) がまとめて HA1 に内部ライト準備応答を送信してもよい (ステップ 2108)。

【0053】

CA (m) 及び CA (b) の両方 (CA (m) が一括して応答する場合は、CA (m)) から内部ライト準備応答を受領した HA1 は、チャンネル 61 を介して

ホストコンピュータ 6 にライト準備応答を送信する。その後、ライト準備応答に応じてホストコンピュータ 6 が送信したライトデータを受領した H A 1 は、内部結合 2 1 を介して、C A (m) 及び C A (b) にライトデータを送信する (ステップ 2 1 1 0)。

【0054】

ライトデータを受領した C A (m) 及び C A (b) は、受信したライトデータをキャッシュメモリ 3 2 内の前述したステップで確保された領域に書き込み、対応するディレクトリ情報 3 6 2 を更新する。その後、C A (m) 及び C A (b) は、内部結合 2 1 を介して、内部ライト要求を送信した H A 1 に対し内部メッセージである内部ライト完了応答を送信する (ステップ 2 1 1 1)。

【0055】

C A (m) 及び C A (b) の両方から内部ライト完了応答を受領した H A 1 は、チャンネル 6 1 を介して、ホストコンピュータ 6 にライト完了応答を送信する (ステップ 2 1 1 3)。

【0056】

以下、4つのキャッシュアダプタ 3 間の操作を例として、キャッシュメモリ 3 2 の領域割り当ての方法について説明する。簡単のため、以下各キャッシュメモリ 3 2 が有する記憶領域のうち、ライトデータを格納するライトキャッシュ領域のみを図示して説明する。しかし、実際には、キャッシュメモリ 3 2 は、リードデータを格納するリードキャッシュ領域なども有する。又、キャッシュアダプタ 3 を区別するため、以下、各キャッシュアダプタを C A 1、2、3 及び C A 4 と称する。

【0057】

図 6 は、障害が発生していない通常時における、キャッシュアダプタ 3 間でのキャッシュメモリ 3 2 における記憶領域の割り当てを示した図である。本図において、C A 1 及び C A 2、C A 3 及び C A 4 がそれぞれディスク装置 4 を共有し、冗長性を確保するためのペアになっている (この関係を「キャッシュアダプタペア」と呼ぶ)。C A 1 のキャッシュメモリ 3 2 は、C A 1 がマスタキャッシュアダプタとしてステージング動作、デステージング動作を行うべきデータが格納

されるライトキャッシュ領域CA1 (M) 30121及びCA2のバックアップキャッシュアダプタとしてライトデータの複製を格納するライトキャッシュ領域CA2 (B) 30122を有する。

【0058】

以下、マスタキャッシュアダプタに扱われるデータが格納される記憶領域をマスタ領域と呼び (M) で表し、バックアップキャッシュアダプタに扱われるデータ (複製されたデータ) が格納される記憶領域をバックアップ領域とよび (B) で表す。

【0059】

CA1のキャッシュアダプタペアであるCA2のキャッシュメモリ32は、ライトキャッシュ領域CA2 (M) 30221、ライトキャッシュ領域CA1 (B) 30222が含まれる。従って、ライトキャッシュ領域CA1 (M) 30121に含まれるデータと、ライトキャッシュ領域CA1 (B) 30222に含まれるデータは同一である (アドレス、並び順は異なってもよい)。

【0060】

つまり、本図においては、キャッシュアダプタペアである二つのキャッシュアダプタ3が、お互いのデータの複製を格納する記憶領域をキャッシュメモリ32に設け、キャッシュアダプタペアのライトデータを冗長化している。本図の矢印で、上述したデータの複製の関係を示している。CA1とCA2のキャッシュアダプタペアと同様に、CA3とCA4のキャッシュアダプタペアも、双方のキャッシュメモリ32に格納されたデータを冗長化している。

【0061】

図7は、CA2が有するキャッシュメモリ32に障害が発生したり除去されたりといった要因で使用不能となった場合の、キャッシュメモリの記憶領域の割り当てを示した図である。ここでは、一方の電源部に障害が発生した場合にも運転が継続できるよう、2つの電源部A511、電源部B512それぞれがCA1とCA3、CA2とCA4に電力を供給している。CA2が有するキャッシュメモリ32が使用不能となった場合、CA1は、CA1が有するCA2のマスタ領域の複製、即ちCA2 (B) を、CA2のマスタ領域として取り扱う。つまり、C

A1がライトキャッシュ領域CA2 (M) 30221に関してもマスタキャッシュアダプタとして動作する。

【0062】

一方、CA2及びCA1に配置されていたバックアップ領域であるライトキャッシュ領域CA1 (B) 30222及びライトキャッシュ領域CA2 (B) 30122は、CA4に配置される。具体的には、CA1のキャッシュメモリ32に格納されたデータの複製をCA4のキャッシュメモリ32に格納する。こうすることにより、1つのキャッシュアダプタ3が使用不能となった場合においても、キャッシュアダプタ3に格納されたライトデータは、キャッシュアダプタペアでは無い他のキャッシュアダプタ3によって冗長化されることになる。

【0063】

尚、CA4の代わりにCA3にCA1のデータの複製を配置することも可能であるが、CA1に電力を供給している電源部A511と別の電源部B512に電力を供給されているCA4にデータの複製を配置することにより、さらにどちらかの電源部に障害が発生した場合においても、マスタ領域、バックアップ領域のいずれかには電力が供給されているのでライトデータを失うことがなくなる。

【0064】

ただし、CA4のキャッシュメモリ32の容量がCA1と同じ場合、CA4のキャッシュメモリ3042にCA1が有するデータ全てを格納することはできない (CA4の通常の使用が不可能になる) ので、CA1は、キャッシュメモリ32に格納されているデータの半分をディスク装置4へデステージングした後、残ったデータの複製をCA4へ配置する。尚、デステージングするデータ量は特に半分でなくても良いが、CA1及びCA4のキャッシュメモリ32のキャッシュヒット率を考慮すると、半分にの方が望ましい。

【0065】

図8は、図7と同じ状況におけるキャッシュメモリの記憶領域の割り当ての別例を示した図である。図7とは、CA1へのデータの格納方法は同じだが、CA1に格納されたデータの複製の配置が異なる。すなわち、CA2が使用不能になる前にCA2とCA1に配置されていたバックアップ領域であるライトキャッシュ

領域CA1 (B) 30222とライトキャッシュ領域CA2 (B) 30122
がそれぞれCA4とCA3に配置される。

【0066】

この場合、CA1がデステージングするデータ量は、図7とは異なり、もとの記憶領域の1/3で良い。なぜなら、図7と異なり、CA1に格納されるデータの複製は、CA4及びCA3のキャッシュメモリ32に分散されて配置されるので、個々のキャッシュアダプタ3でCA1のデータの複製が占める領域が減るからである。このことより、1つのキャッシュメモリ32が使用不能となった場合においてもライトデータはキャッシュメモリ32によって冗長化されることになり、図7の場合と比較しライトキャッシュ領域の大きさは大きくなる。

【0067】

図11は、キャッシュメモリの記憶領域の割り当ての更なる別例を示した図である。本例では、更にCA5及びCA6の新たなキャッシュアダプタペアを追加することで、あるキャッシュアダプタ3が使用不可能となった際における1つのキャッシュアダプタ3当たりのライトキャッシュ領域の大きさを、CA1のバックアップデータを格納しつつも他の例と比較して大きくすることができる。図11は、4つのキャッシュアダプタ3にバックアップ領域を置いた場合のキャッシュメモリの領域割り当て配置例を示している。しかし、バックアップ領域が割り当てられるキャッシュアダプタの数に制限は無い。

【0068】

図9の(A)は、管理アダプタ7が保持するマスタ管理テーブル71の一例を示す図である。マスタ管理テーブル71は、キャッシュアダプタ対応テーブル711及びキャッシュアダプタペアテーブル712を含んでいる。キャッシュアダプタ対応テーブル711は、複数のエントリを有する。各エントリは、LDEV番号、マスタキャッシュアダプタ、バックアップキャッシュアダプタの情報が登録されるフィールド7111、7112及び7113を有する。

【0069】

尚、マスタ管理テーブル71には、ディスク制御装置5全体のLDEVに関する情報が登録される。一方、キャッシュアダプタ3が保持する管理テーブル31

には、そのキャッシュアダプタ 3 がマスタキャッシュアダプタ及びバックアップキャッシュアダプタとして処理すべき L D E V に関する情報のみが登録される。

【0070】

キャッシュアダプタペアテーブル 712 は、複数のエントリを有する。各エントリは、キャッシュアダプタペア及び障害サポートキャッシュアダプタの情報が登録されるフィールド 7121 及び 7122 を含んでいる。以下、キャッシュアダプタペアを括弧を用いて表す。

【0071】

障害サポートキャッシュアダプタとは、同一エントリの対応するキャッシュアダプタペアの一方のキャッシュアダプタ 3 に障害が発生した場合に、バックアップ領域の格納を担当するキャッシュアダプタ 3 である。例えば、図 9 (A) のキャッシュアダプタペアテーブル 712 の第一のエントリでは、CA2 に障害が発生した場合、CA3 及び CA4 に CA1 及び CA2 のバックアップ領域が設けられ、結果としてキャッシュメモリの領域割り当て配置が図 8 のようになることを示している。

【0072】

図 9 の (B) は、図 9 (A) の状態から CA2 に障害が発生した場合に、キャッシュアダプタペアテーブル 712 に基づき内容が変更されたキャッシュアダプタ対応テーブル 711 を示している。テーブル中の網掛けの欄が変更された部分である。マスタキャッシュアダプタの情報が登録されるフィールド 7112 のうち、障害が発生したキャッシュアダプタ 3 である CA2 を指定する情報が登録されていたフィールドの情報がキャッシュアダプタペアの CA1 に、バックアップキャッシュアダプタの情報が登録されるフィールド 7113 のうち、CA2 及び CA1 を指定する情報が登録されていた部分が、キャッシュアダプタペアテーブル 712 に登録された障害サポートキャッシュアダプタの情報に従って、CA3 および CA4 に変更される。

【0073】

図 10 は、CA2 に障害が発生した場合の管理アダプタ 7 (以下「MA」と記述) の処理手順を示す図である。

まずMAはCA2の障害発生を認識する。具体的には、CA2のキャッシュメモリ3022の障害発生報告、CA2に内部メッセージを送信して障害のため応答が得られなかったホストアダプタ1の報告、又はMAの定期的なディスク制御装置5全体の調査やMAに接続された管理インタフェース（図示せず）を介した管理者の指示などにより、MAはCA2のキャッシュメモリ32の障害発生を認識する（ステップ2201）。

【0074】

CA2の障害を認識したMAは、マスタ管理テーブル71のキャッシュアダプタペアテーブル712より、CA2のキャッシュアダプタペアがCA1であることを確認し（ステップ2202）、キャッシュアダプタ対応テーブル711のフィールド7112を走査して、マスタキャッシュアダプタとしてCA2が指定されている部分をキャッシュアダプタペアであるCA1に変更する（ステップ2203）。

【0075】

続いて、MAは、キャッシュアダプタペアテーブル712よりキャッシュアダプタペアである（CA1、CA2）の障害サポートキャッシュアダプタが（CA3、CA4）であることを確認する（ステップ2204）。その後、MAはキャッシュアダプタ対応テーブル711のフィールド7113を走査し、CA1及びCA2を指定している部分の内容を障害サポートキャッシュアダプタであるCA3又はCA4に変更する（ステップ2205）。

【0076】

その後、MAは、キャッシュアダプタ対応テーブル711で内容が変更されたエントリの情報を、ホストアダプタ1及び変更されたエントリの変更前または変更後のフィールド7112及び7113に登録されていたキャッシュアダプタ3に配信する（ステップ2206）。

【0077】

この配信された情報を受信したホストアダプタ1又はキャッシュアダプタ3は、配信された情報を、自身が有する管理テーブル11又は管理テーブル31に反映する。さらにキャッシュアダプタ3は、キャッシュメモリ3のライトキャッシ

ュ領域割り当てを計算する(ステップ2207)。尚、ステップ2207の処理については後に詳述する。このようにして、MAがキャッシュアダプタ3の障害を検出し、バックアップキャッシュアダプタの設定を変更してシステム全体に変更した情報を配信する。これにより、ホストアダプタ1はマスタキャッシュアダプタとしてCA2を使用していたLUへのアクセスをCA1へ変更することができ、CA1は、CA2が担当していたディスク装置4に対するステージング動作及びデステージング動作を引継ぐことができる。

【0078】

図12は、CA2に障害が発生した際に、MAから配信された情報を受信したキャッシュアダプタ3における処理手順を示した図である。

まず、CA2のキャッシュアダプタペアであるCA1は、CA2又は内部メッセージを送信したホストアダプタ1の報告、もしくはMAの調査やMAを介した管理者の指示などにより、CA2のキャッシュメモリ3022の障害発生を認識する(ステップ2301)。

【0079】

障害発生を認識したCA1は、CA1及びCA2が管理するLDEVに対するアクセス要求受領を中止し、キャッシュメモリ3012のライトキャッシュ領域に格納されているデータをディスク装置4に書き込む(ステップ2302)。その後、更新されたキャッシュアダプタ対応テーブル711のエントリの情報の一部(CA1に関係する部分のみ)をMAより受領し、管理テーブル31に反映する(ステップ2303)。

【0080】

CA1は、反映された内容に基づき、キャッシュメモリ3のライトキャッシュ領域割り当てを計算する。具体的には、エントリ7122に登録されている障害サポートキャッシュアダプタの数をmとすると、障害発生により、データのバックアップに使用される領域以外の1つのキャッシュアダプタ3で利用できるライトキャッシュ領域は、ライトキャッシュ領域全体の $m / (2m + 2)$ となる(ステップ2304)。

【0081】

計算の結果に基づいてCA1は、バックアップキャッシュアダプタとなる障害サポートキャッシュアダプタ、ここではCA3及びCA4に、内部メッセージであるライトキャッシュ領域割当要求を送信する(ステップ2305)。ライトキャッシュ領域割当要求を受信したCA3及びCA4は、要求されたCA1およびCA2のデータをバックアップするライトキャッシュ領域が確保できるまで、キャッシュメモリ3のデータをディスク装置4に書き込む(ステップ2306)。

【0082】

CA1に要求されたライトキャッシュ領域を確保したCA3及びCA4は、内部メッセージであるライトキャッシュ領域割当返答をCA1に送信する(ステップ2307)。CA1は、ライトキャッシュ領域割当要求を送信した全てのキャッシュアダプタ3からのライトキャッシュ領域割当返答を確認し、ステップ2302でアクセス受領を中止したLDEVに対するアクセス要求受領を再開する(ステップ2308)。

【0083】

この結果、CA1のキャッシュメモリに書き込まれたデータはCA3又はCA4にバックアップされ、冗長性が確保される。尚、データの書き込みは上述した処理手順で行われるが、ライトデータのバックアップの対象がCA2では無く、CA3又はCA4となる。

【0084】

次に、障害が発生したキャッシュアダプタ3のキャッシュメモリ32が回復した場合の処理を説明する。図13は、CA2に発生した障害が回復した時の処理手順を示した図である。

MAは、管理者の指示などからCA2の回復を認識し(ステップ2401)、キャッシュアダプタ対応テーブル711をCA2に障害が発生する以前の状態に変更し、内部メッセージにより各ホストアダプタ1及び各キャッシュアダプタ3にその変更された情報を配信する。尚、キャッシュアダプタ対応テーブル711の障害発生前の状態は、MAが有するメモリに格納されている。またMAは、CA2のキャッシュアダプタペアであるCA1に、内部メッセージを用いてCA2の障害回復を通知する(ステップ2402)。

【0085】

通知を受けたCA1は、CA1のライトキャッシュ領域に格納された全てのデータをディスク装置4に書込み、ライトキャッシュ領域のデータを無効化する（ステップ2403）その後、CA1は、バックアップキャッシュアダプタの動作を行っていた障害サポートキャッシュアダプタであるCA3及びCA4に内部メッセージであるライトキャッシュ領域開放要求を送信する（ステップ2404）。

【0086】

ライトキャッシュ領域開放要求を受信したCA3及びCA4は、CA1及びCA2のバックアップ領域に該当するライトキャッシュ領域を開放し、ライトキャッシュ領域をCA2に障害が発生する以前のCA3とCA4各々のマスタ領域、バックアップ領域に変更し、CA1にライトキャッシュ領域開放返答を送信する。尚、障害が発生する以前のCA3及びCA4のマスタ領域及びバックアップ領域の情報はMAに保存されており、CA3及びCA4はMAと内部メッセージを用いてつうしんすることにより、これらの情報を取得して、キャッシュメモリ32に構成を変更する（ステップ2405）。

【0087】

CA3及びCA4からのライトキャッシュ領域開放返答を確認したCA1は、CA2に内部メッセージである動作開始要求を送信し、CA1の担当するLDEVに対するアクセス要求受領を開始する（ステップ2406）。動作開始要求を受けたCA2は、CA2がマスタキャッシュアダプタとなるLDEVに対するアクセス要求受領およびバックアップキャッシュアダプタとしての処理を開始する（ステップ2407）。

【0088】

次に、キャッシュアダプタペアの双方のキャッシュアダプタのキャッシュメモリ32に障害が発生した場合においてもライトデータを失わず障害から回復する処理を説明する。

図14は、キャッシュアダプタペアであるCA1及びCA2に障害が発生し、キャッシュアダプタ3の交換などにより障害から回復するまでの処理手順を示し

た図である。尚、本実施形態では、CA1又はCA2のいずれか一方に障害が発生してCA3及びCA4にCA1及びCA2のキャッシュメモリ32に保存されるべきデータがバックアップされている状態で、残りのCAに障害が発生したとして説明する。

【0089】

まず、キャッシュアダプタペアであるCA1及びCA2の双方のキャッシュメモリ32に障害が発生する（ステップ2501）。

MAは、CA1、CA2又はホストアダプタ1の報告、MAの調査若しくは管理者の指示などにより、キャッシュアダプタペアであるCA1及びCA2双方のキャッシュメモリ32の障害発生を認識する（ステップ2502）。

【0090】

この場合MAは、ホストアダプタ1に、CA1及びCA2がマスタキャッシュアダプタとなるLDEVの使用不可要求を送信する。使用不可要求を受信したホストアダプタ1は、該当LDEVへのアクセスを要求するホストコンピュータ6にアクセス不能のエラーを返す（ステップ2503）。この状態から保守員により、CA1及びCA2が新しいキャッシュアダプタ3に交換され、元のようにディスク装置4がディスク側チャネル41に接続され、新しいキャッシュアダプタ3がディスク制御装置5全体からCA1、CA2として認識されるように設定され、CA1及びCA2のキャッシュメモリ32は障害から回復する（ステップ2504）。

【0091】

障害から回復したCA1及びCA2は、それぞれ、バックアップキャッシュアダプタ動作を行っていた障害サポートキャッシュアダプタであるCA3及びCA4に内部メッセージであるライトデータ送信要求を送信する（ステップ2505）。ライトデータ送信要求を受信したCA3及びCA4は、それぞれ、CA1、CA2のバックアップ領域に該当するライトキャッシュ領域に格納されていたデータをCA1又はCA2に送信する。該当するライトキャッシュ領域に格納されていたデータを全て送信後、CA3及びCA4は、該当するライトキャッシュ領域を開放し、ライトキャッシュ領域をCA1、CA2に障害が発生する以前のC

A3とCA4各々のマスタ領域、バックアップ領域に変更する（ステップ2506）。

【0092】

CA1及びCA2は、CA3又はCA4より受信したライトデータを逐次ディスク装置4に書き込み、全ライトデータを処理後、各CAが担当するLDEVに対するアクセス要求の受領を開始する（ステップ2507）。

【0093】

ここまではMAの主導によるキャッシュアダプタ3の記憶領域割り当ての処理を説明した。しかし、各ホストアダプタ1及び各キャッシュアダプタ3だけで上述した処理を行うことも可能である。

【0094】

図15は、CA2のキャッシュメモリ32に障害が発生した場合のキャッシュアダプタ3の記憶領域割り当て処理で、MAを使用しない実施形態を示した図である。なおこの場合は、各ホストアダプタ1及び各キャッシュアダプタ3は、各自がアクセスしうるキャッシュアダプタ3に対応したキャッシュアダプタ対応テーブル711及びキャッシュアダプタペアテーブル712を、各自の管理テーブル11および管理テーブル31に所持している。

【0095】

また、これまでの説明においては、キャッシュメモリ32の障害が発生した時点で、まずキャッシュアダプタペアの他方のキャッシュアダプタ3のキャッシュメモリ32に格納されているライトデータを全てディスク装置4に書き込む処理方式を述べてきたが、全てのライトデータをディスク装置4に書き込まない方式をとってもよい。ここで、前者を全デステージ方式、後者をコピー方式と呼び、本実施形態では、両方式の処理について説明する。

【0096】

なお、コピー方式はこれまで説明してきたMAを介する処理においても同様に実行できる。両方式の選択は、処理時間と信頼性の兼ね合いで決定される。以下、図16に示した処理手順についての説明を行う。

【0097】

ホストアダプタ 1 (以下「HA 1」) は、内部リード要求、内部ライト要求の応答などから CA 2 のキャッシュメモリ 3 2 の障害を認識し、他の全 HA 1 に内部メッセージにより CA 2 の障害を通知する (ステップ 2601)。通知を受けた全 HA 1 は、各々が有するキャッシュアダプタペアテーブル 7 1 2 に基づいて、CA 2 のキャッシュアダプタペアが CA 1 であること及びその障害サポートキャッシュアダプタを確認し、管理テーブル 1 1 のフィールド 1 1 2 2 が CA 2 であるものを CA 1 に、フィールド 1 1 2 3 が CA 2 又は CA 1 であるものをその障害サポートキャッシュアダプタに変更する (ステップ 2602)。さらに、障害を発見した HA 1 は、CA 1 に内部メッセージにより CA 2 の障害を通知する (ステップ 2603)。

【0098】

通知を受けた CA 1 は、設定されている方式に応じて、以下の処理を行う。全デステージ方式の場合、CA 1 は、ライトキャッシュ領域に格納されている全てのデータをディスク装置 4 に書き込み、そのライトキャッシュ領域を無効化する。次に、障害サポートキャッシュアダプタの数に基づいてライトキャッシュ領域割り当てを計算し、管理テーブル 3 1 を変更する。一方、コピー方式の場合は、CA 1 はライトキャッシュ領域の無効化を行わずにライトキャッシュ領域割り当てを計算し、計算したライトキャッシュ領域が確保できるだけディスク装置 4 にライトデータを書き込み、該当ライトキャッシュ領域を無効化する (ステップ 2604)。

【0099】

続いて CA 1 は、バックアップキャッシュアダプタとなる障害サポートキャッシュアダプタに計算した分のライトキャッシュ領域割当要求を内部メッセージにより送信する (ステップ 2605)。ライトキャッシュ領域割当要求を受信した障害サポートキャッシュアダプタは、要求されたライトキャッシュ領域が確保できるだけ、キャッシュメモリ 3 のライトデータをディスク装置 4 に書き込み、キャッシュ領域を無効化する (ステップ 2606)。

【0100】

尚、コピー方式の場合、CA 1 はここでライトキャッシュ領域に格納されたデ

ータを障害サポートキャッシュアダプタに送信する。送信が完了した場合、又は全デステージ方式の場合は、全てのキャッシュアダプタで該当LDEVに対するアクセス要求の受領を開始する（ステップ2607）。

【0101】

以上のように本発明においては、同一のディスク装置を共有する、キャッシュメモリを備えた制御装置（キャッシュアダプタ）のペアをネットワークを介して接続することでストレージシステムを大規模構成とし、制御装置のペア間で、一方の制御装置のキャッシュメモリに相手の制御装置のキャッシュメモリのライトデータの複製を保持することで相互に冗長化し信頼性を向上させる。

【0102】

また、キャッシュメモリ障害発生時には障害が発生した制御装置とディスク装置を共有する制御装置がステージング及びデステージングを行い、ライトデータの冗長化のみを他の正常動作している制御装置に行わせることで、障害発生以前のライトアクセス応答速度と信頼性を維持する。

【0103】

【発明の効果】

本発明によって、キャッシュメモリを有する記憶装置システムにおいて信頼性が向上する。また、キャッシュメモリを有する記憶装置システムにおいて、障害発生以前のライトアクセス応答速度と信頼性を維持する。

【図面の簡単な説明】

【図1】

本発明の第1の実施の形態のストレージシステムの概要を表した図である。

【図2】

キャッシュアダプタ3の構成例を表した図である。

【図3】

管理テーブル11および31を表した図である。

【図4】

リード要求の処理の流れを示したフローチャートである。

【図5】

ライト要求の処理の流れを示したフローチャートである。

【図 6】

通常時のキャッシュメモリの領域割り当て配置を示した図である。

【図 7】

CA 2 に障害が発生した場合のキャッシュメモリの領域割り当て配置を示した図である。

【図 8】

CA 2 に障害が発生した場合のキャッシュメモリの領域割り当て配置を示した図である。

【図 9】

マスタ管理テーブル 7 1 を示す図である。

【図 1 0】

CA 2 に障害が発生した場合の管理アダプタ 7 の処理を含めたフローチャートである。

【図 1 1】

ライトキャッシュ領域の大きさの比較を表す図である。

【図 1 2】

CA 2 に障害が発生した場合の他のキャッシュアダプタの処理のフローチャートである。

【図 1 3】

CA 2 に発生した障害が回復した時の処理のフローチャートである。

【図 1 4】

キャッシュアダプタペアに障害が発生し、障害から回復する場合の処理を示したフローチャートである。

【図 1 5】

CA 2 に障害の発生した場合の管理アダプタ 7 を介さない処理を示したフローチャートである。

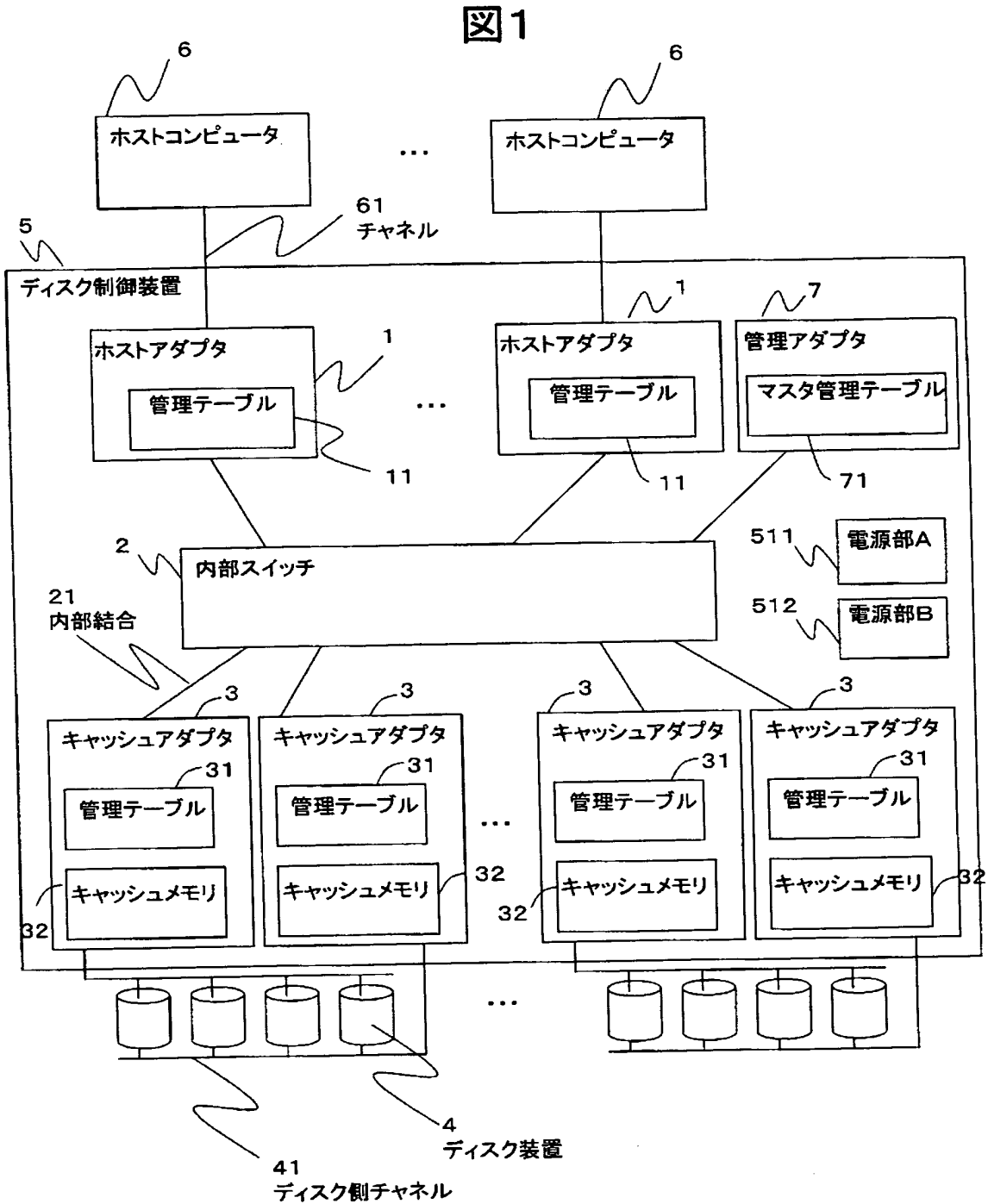
【符号の説明】

1…ホストアダプタ、2…内部スイッチ、3…キャッシュアダプタ、4…ディスク

ク装置、5…ディスク制御装置、6…ホストコンピュータ、7…管理アダプタ、
2 1…内部結合、3 2…キャッシュメモリ、3 3…内部結合I/F部、3 4…ディスク側チャンネルI/F部、3 5…プロセッサ周辺制御部、3 6…制御メモリ、3 8…キャッシュデータバス、3 9…制御データバス、4 1…ディスク側チャンネル、
6 1…チャンネル、5 1 1…電源部A、5 1 2…電源部B。

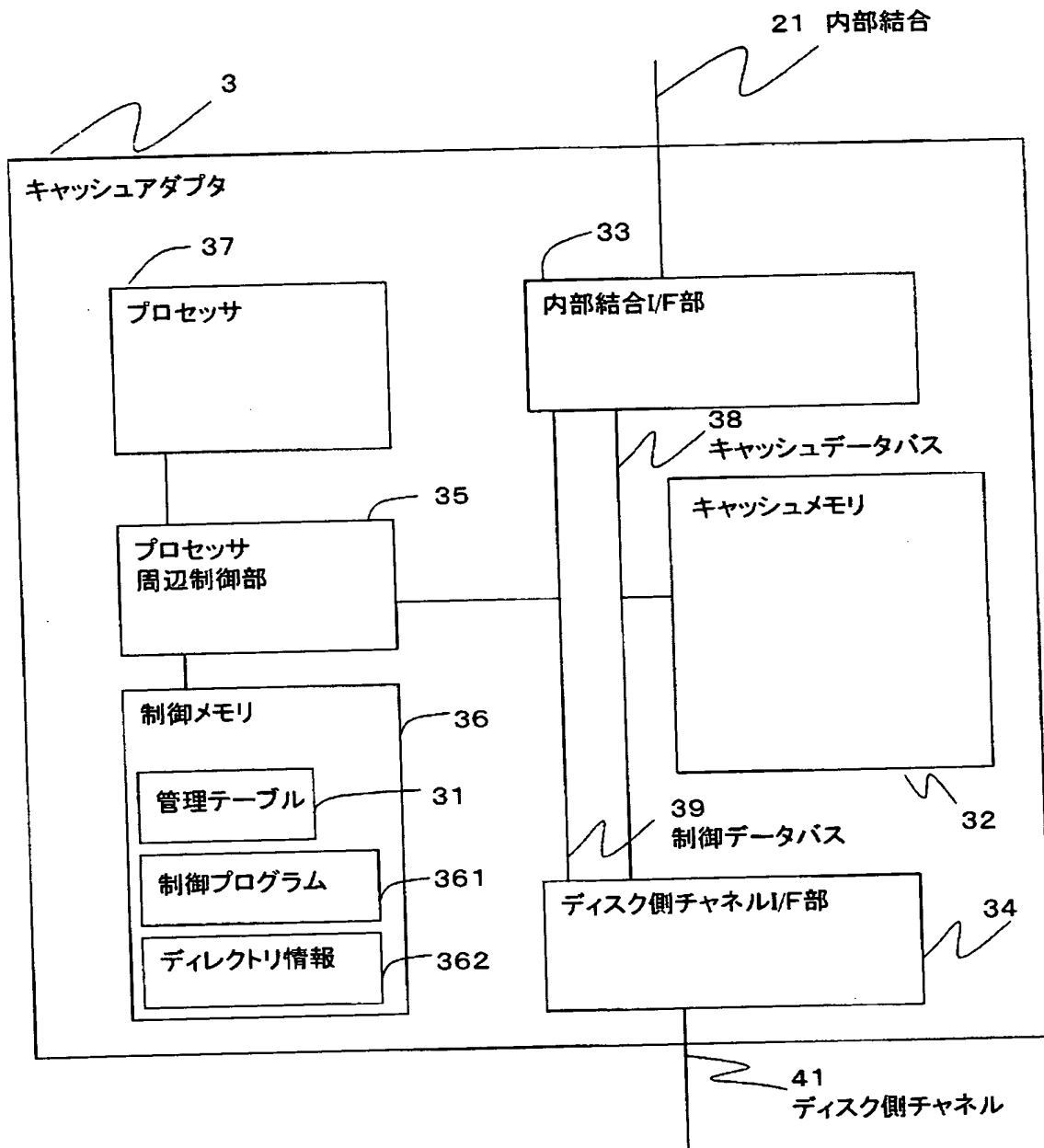
【書類名】 図面

【図 1】



【図 2】

図 2



【図3】

図3

11 管理テーブル
(ホストアダプタ1保持)

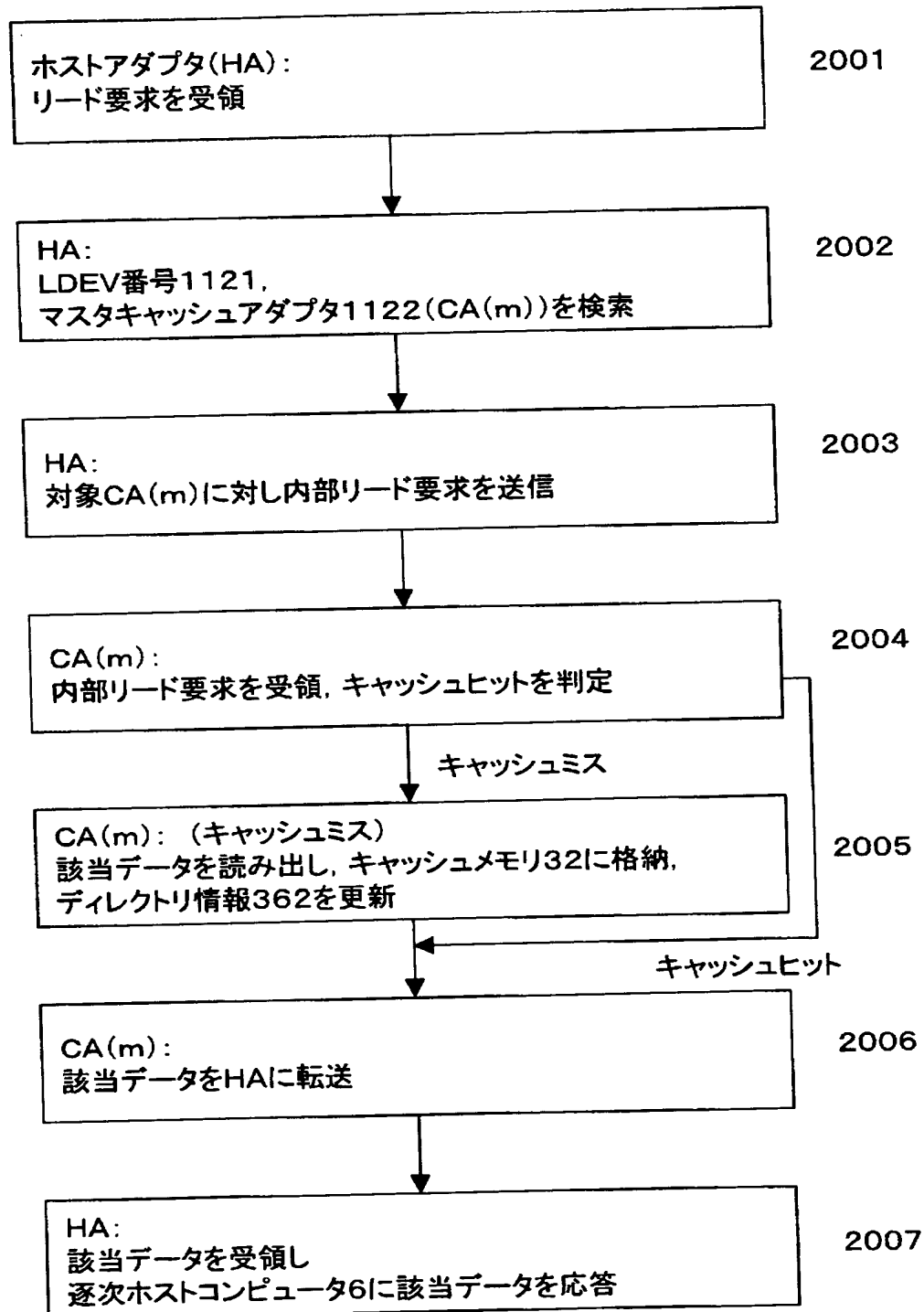
111 アクセス元 指定情報	宛先情報			112
1111 LU番号	LDEV番号	1121 マスタ キャッシュアダプタ	1122 バックアップ キャッシュアダプタ	1123
1	11	CA1	CA2	
2	21	CA2	CA1	
3	41	CA4	CA3	

31 管理テーブル
(キャッシュアダプタ3保持)

311 LDEV番号	312 バックアップ キャッシュアダプタ	313 VDEV番号	314 VDEVアドレス
11	CA2	1	00000000 ~ 002FFFFFF
12	CA2	1	00300000 ~ 008FFFFFF
13	CA2	2	00000000 ~ 006FFFFFF
21	CA1	3	00200000 ~ 007FFFFFF
22	CA1	3	00800000 ~ 0087FFFF

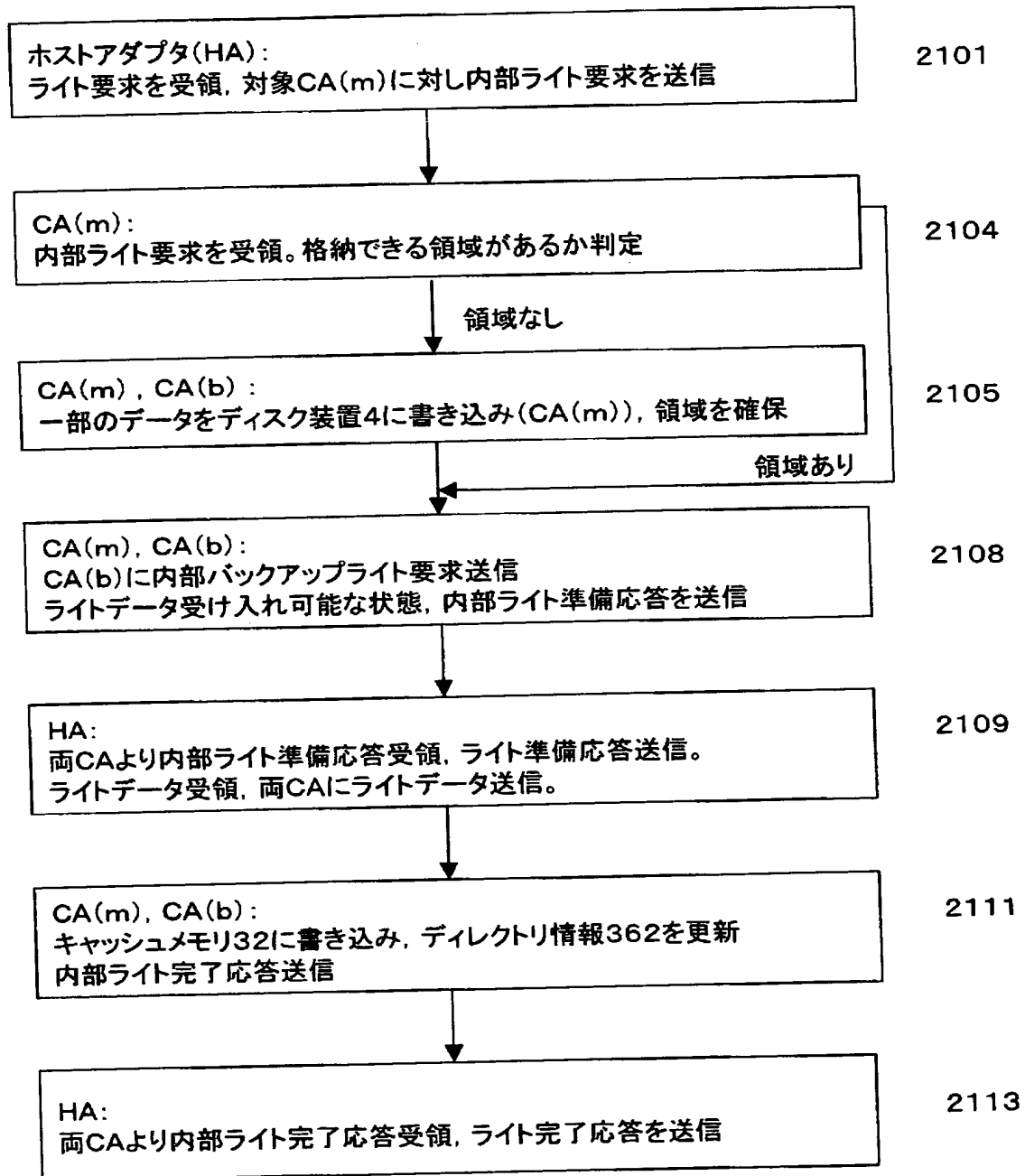
【図 4】

図 4



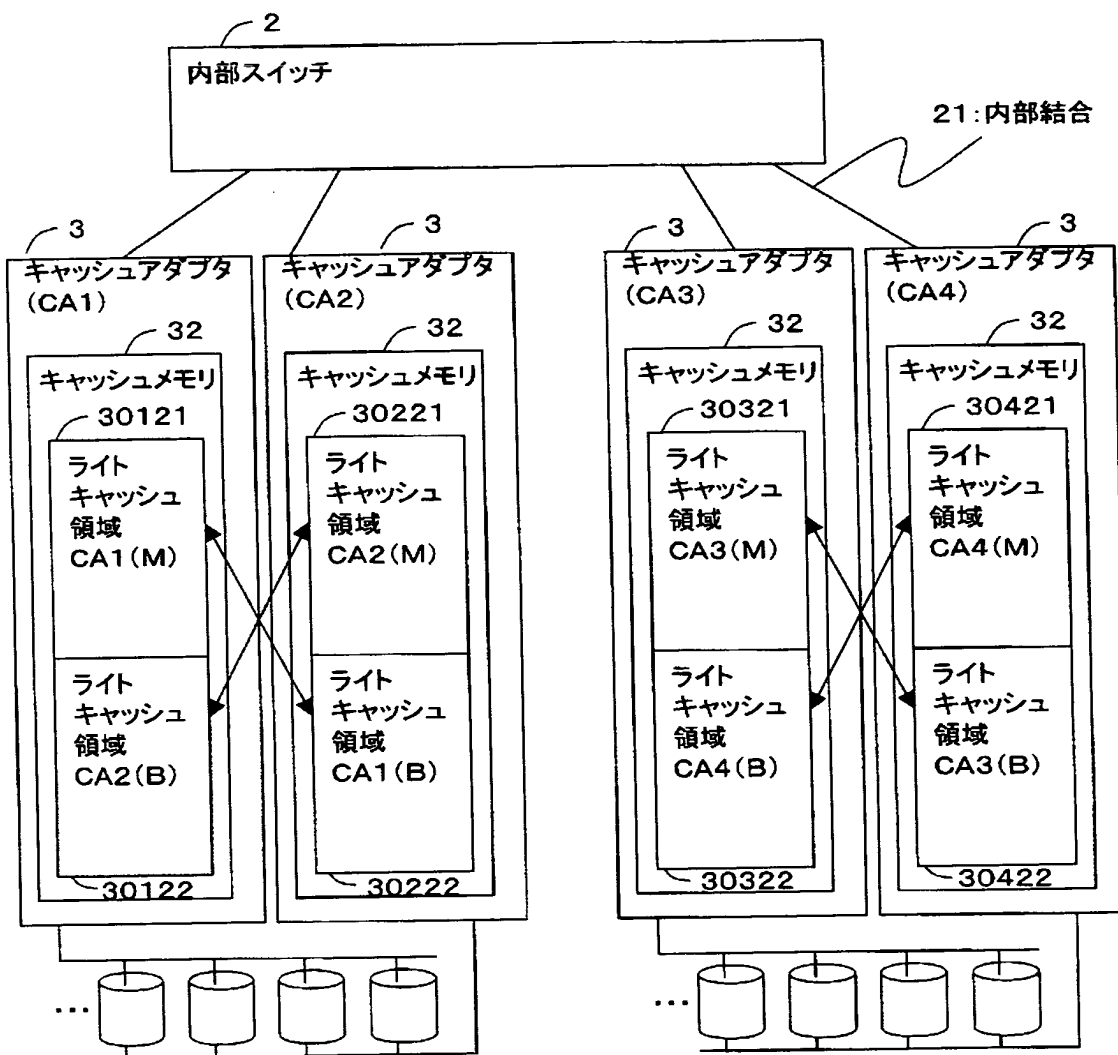
【図5】

図5



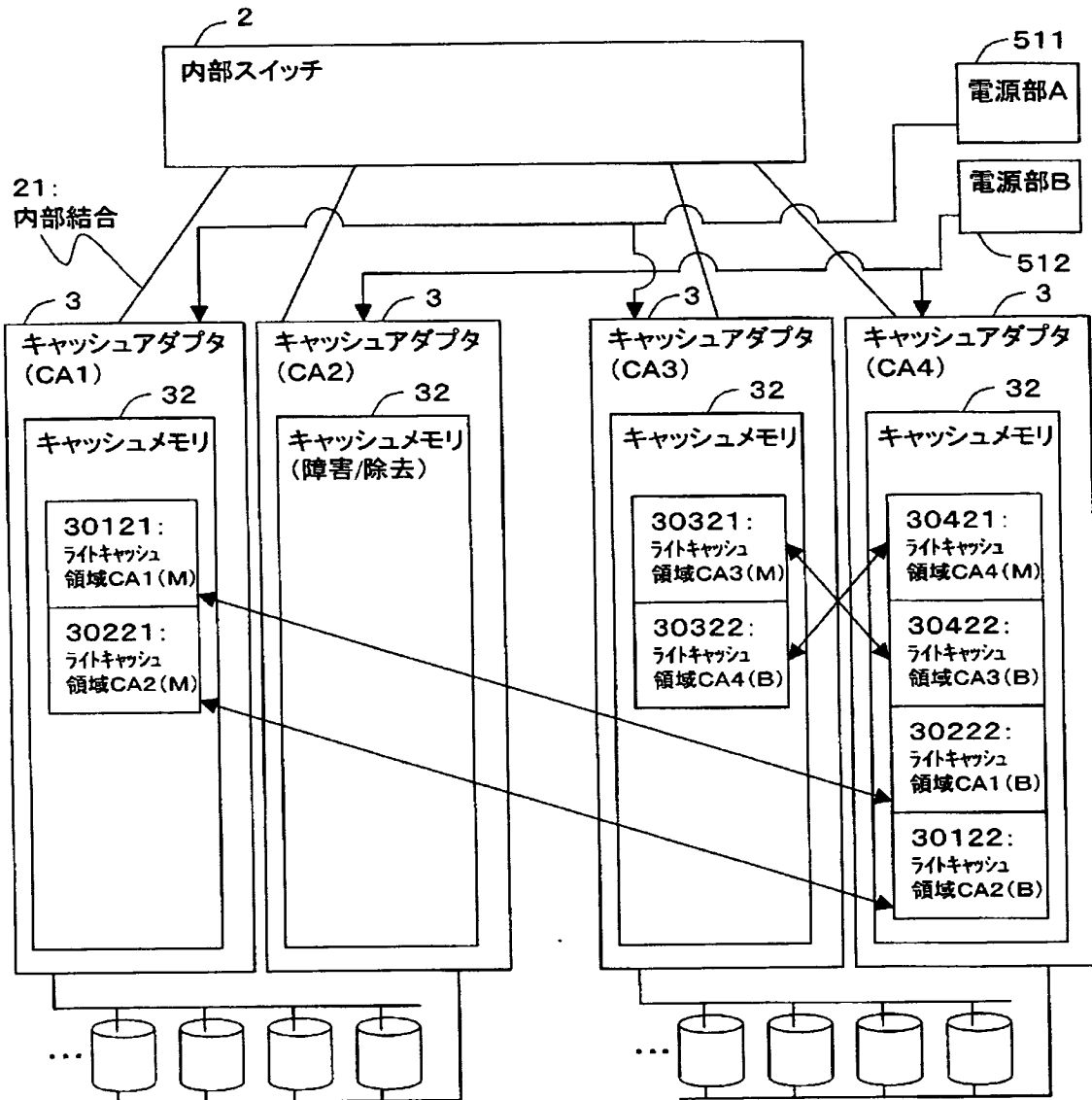
【図 6】

図6



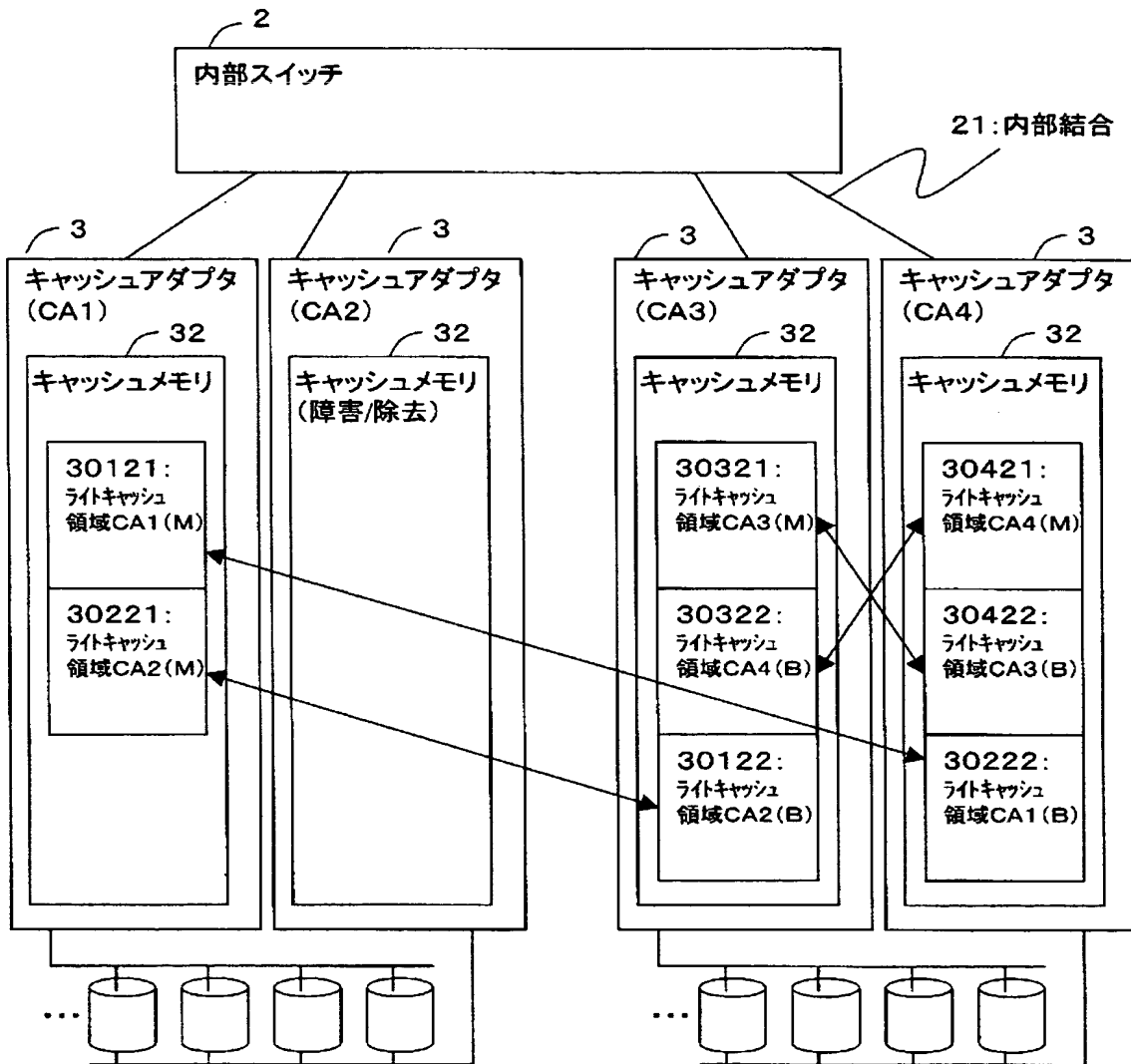
【図 7】

図 7



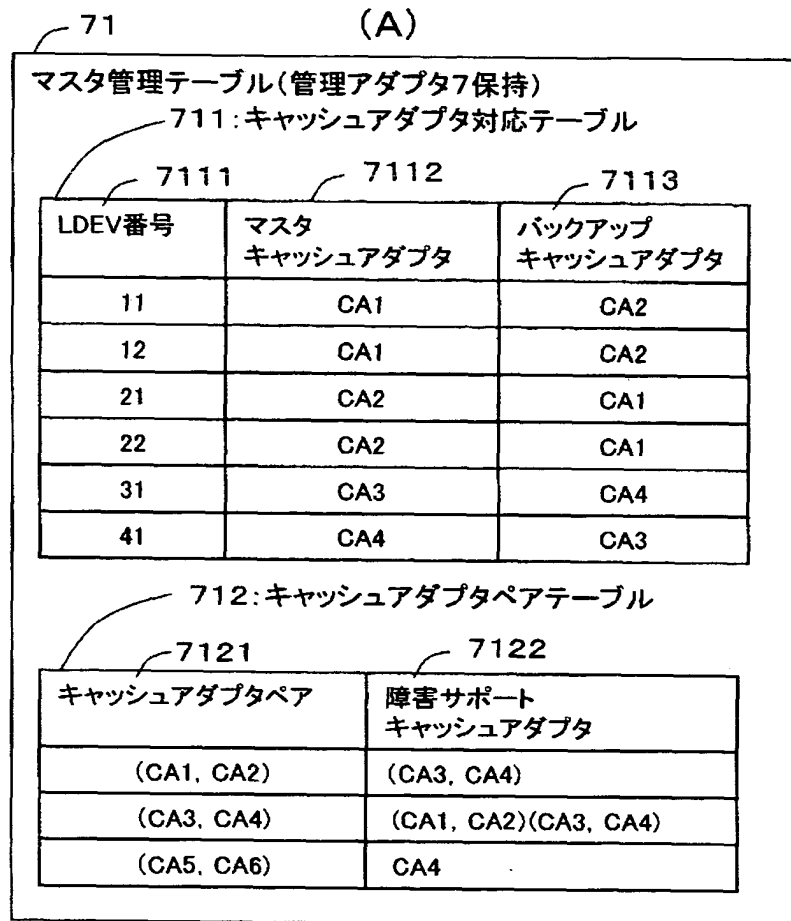
【図 8】

図8



【図9】

図9

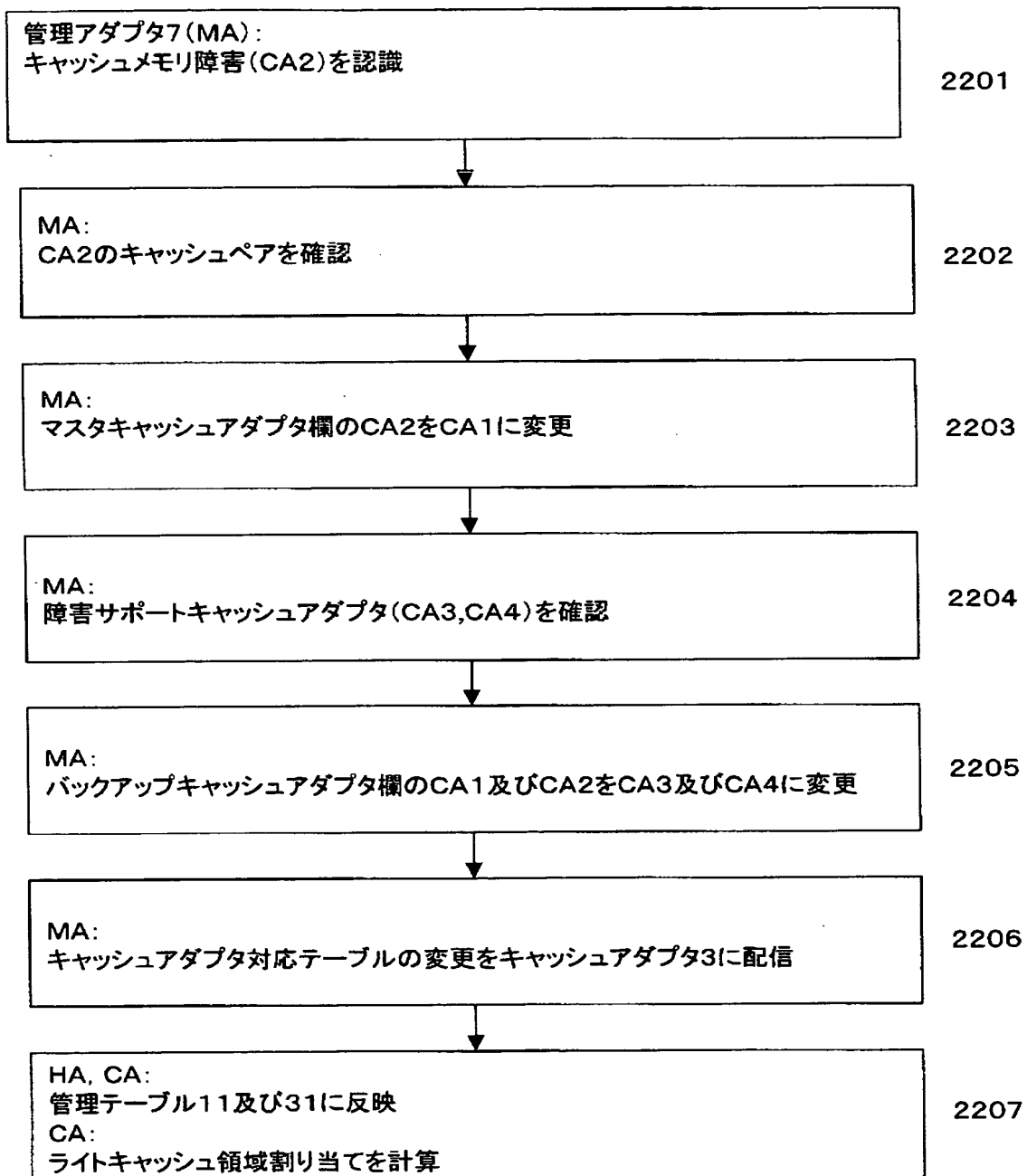


(B)

7111 LDEV番号	7112 マスタ キャッシュアダプタ	7113 バックアップ キャッシュアダプタ
11	CA1	CA4
12	CA1	CA4
21	CA1	CA3
22	CA1	CA3
31	CA3	CA4
41	CA4	CA3

【図10】

図10



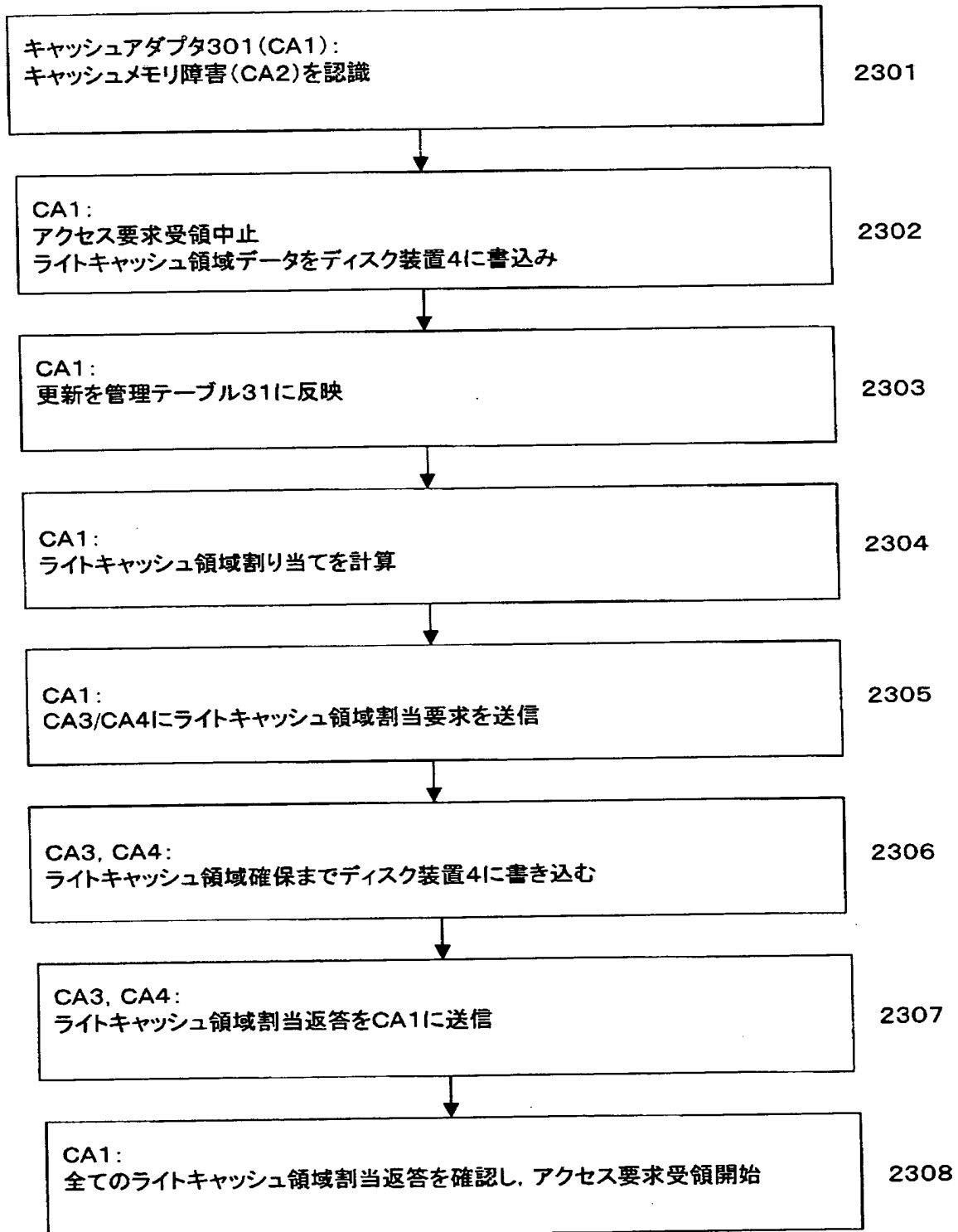
【図 11】

図 11

	32: キャッシュメモリ(CA1)	32: キャッシュメモリ(CA2)	32: キャッシュメモリ(CA3)	32: キャッシュメモリ(CA4)	32: キャッシュメモリ(CA5)	32: キャッシュメモリ(CA6)
0	30121: ライトキャッシュ 領域CA1(M)		30321: ライトキャッシュ 領域CA3(M)	30421: ライトキャッシュ 領域CA4(M)	30521: ライトキャッシュ 領域CA5(M)	30621: ライトキャッシュ 領域CA6(M)
1						
2	30221: ライトキャッシュ 領域CA2(M)		30322: ライトキャッシュ 領域CA4(B)	30422: ライトキャッシュ 領域CA3(B)	30522: ライトキャッシュ 領域CA6(B)	30622: ライトキャッシュ 領域CA5(B)
3						
4			301221: ライトキャッシュ 領域CA2(B)	301222: ライトキャッシュ 領域CA2(B)	302221: ライトキャッシュ 領域CA1(B)	302222: ライトキャッシュ 領域CA1(B)
5						

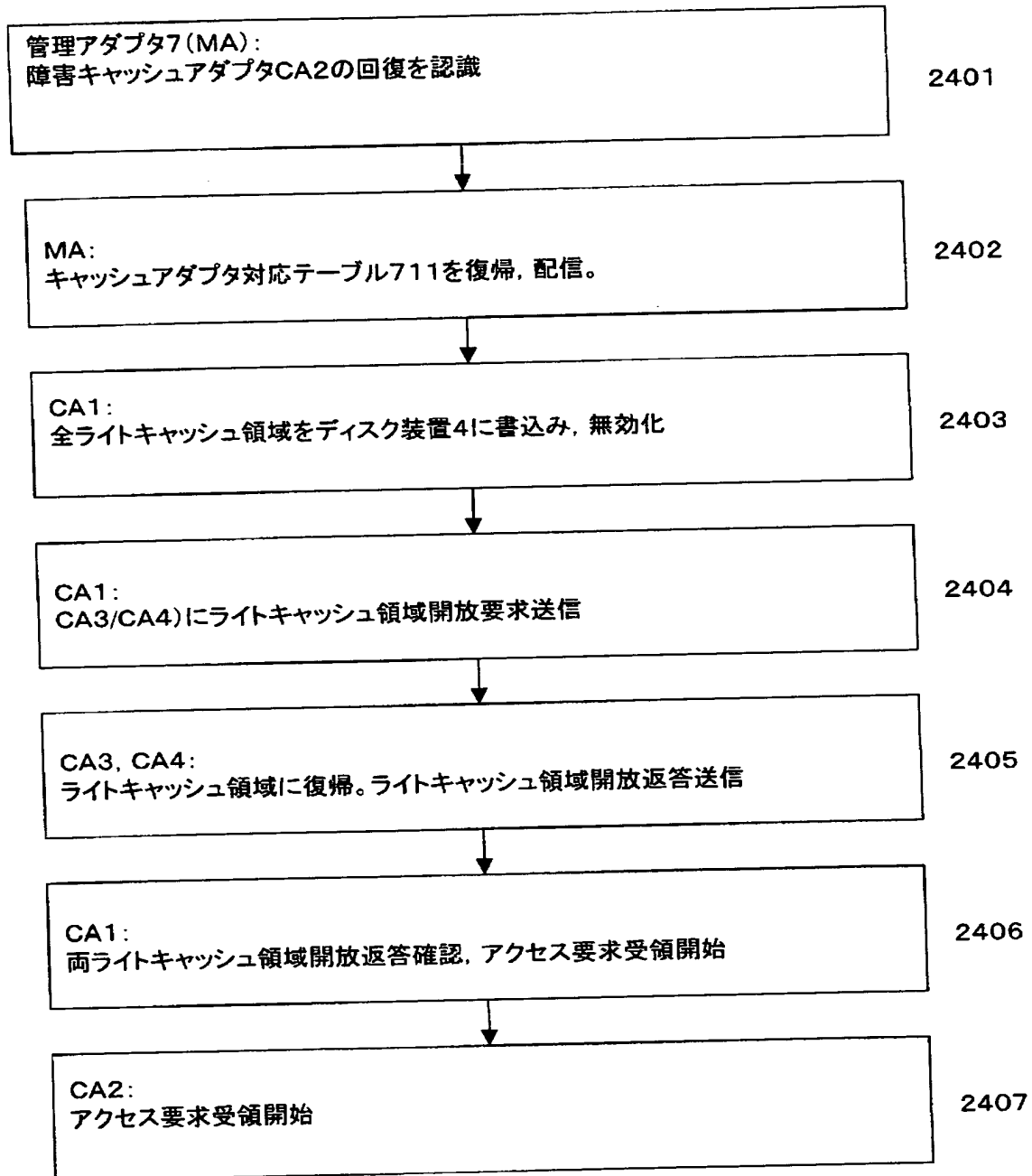
【図 12】

図 12



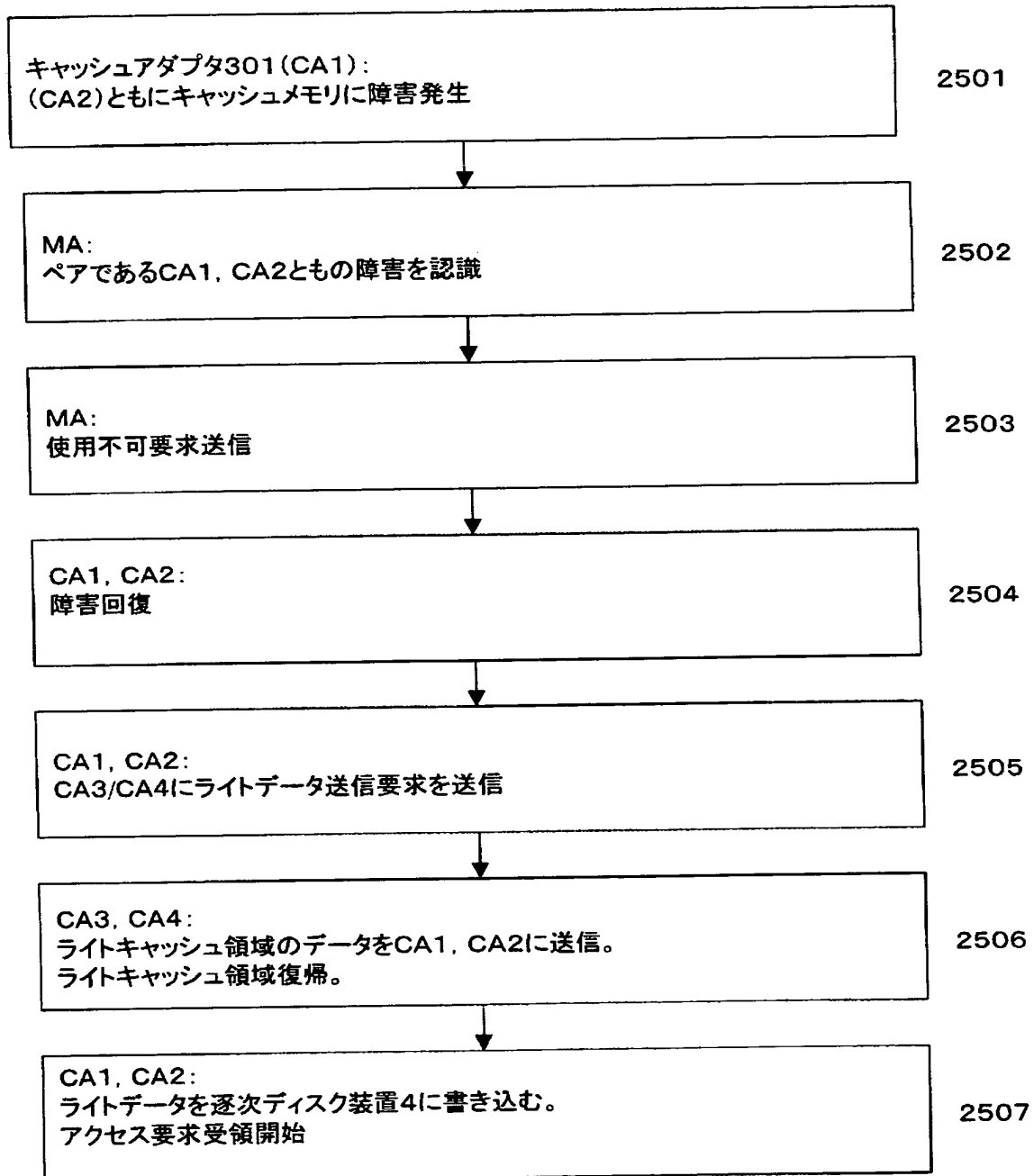
【図13】

図13



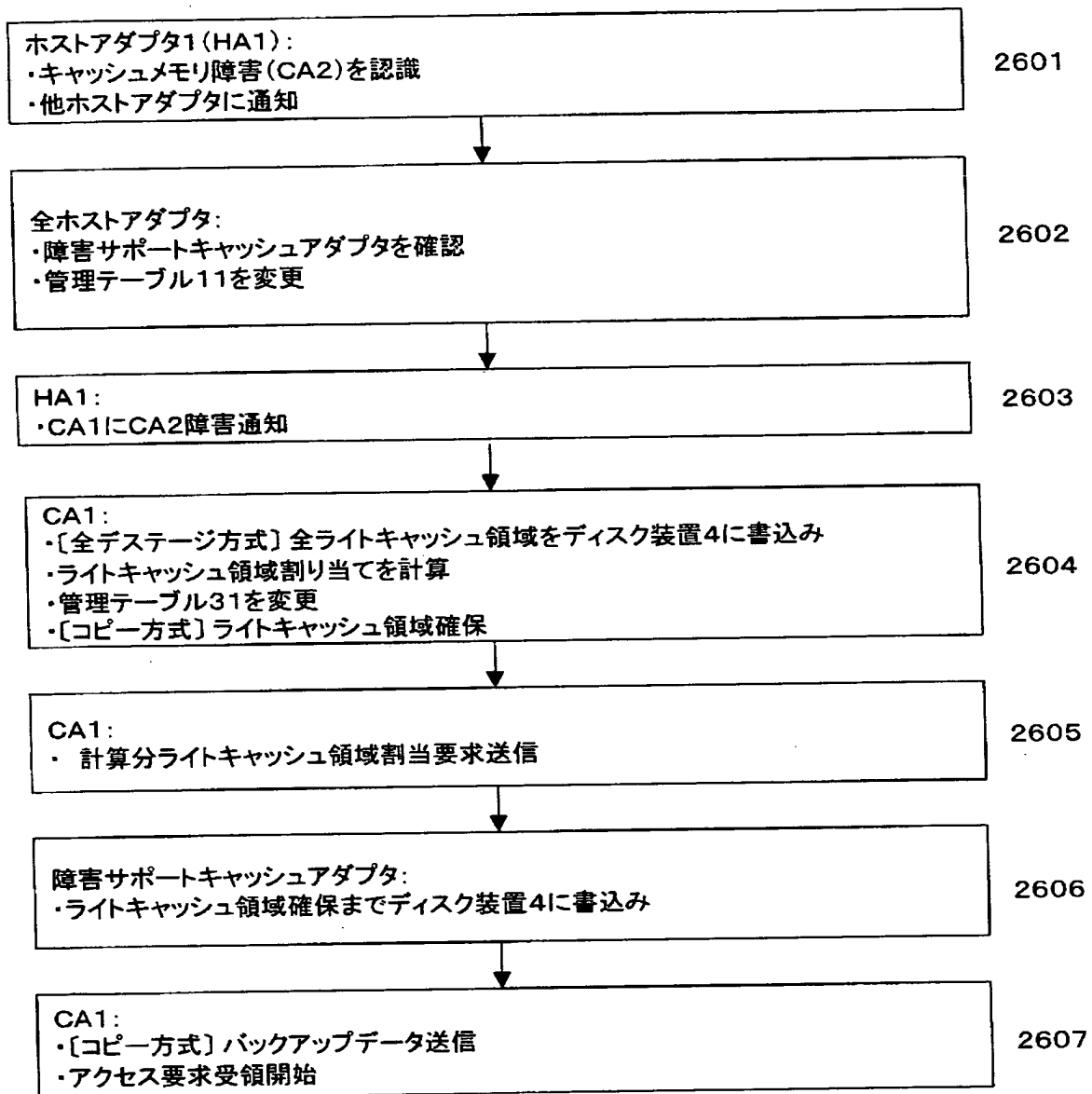
【図 14】

図 14



【図15】

図15



【書類名】 要約書**【要約】****【課題】**

キャッシュ障害発生時においてもライトアクセス応答速度と信頼性を維持する、大規模構成可能なストレージシステムおよびその制御方法を提供する。

【解決手段】

接続されたディスク装置を制御しキャッシュメモリを備えた複数のクラスタを相互結合網によりそれぞれ接続し、同一のディスク装置を共有する2つのクラスタのキャッシュメモリに相手のクラスタのキャッシュメモリのライトデータを相互に冗長化する。

キャッシュメモリ障害発生時には、障害が発生したキャッシュメモリを有するクラスタと同一のディスク装置を共有するクラスタが、障害が発生したキャッシュメモリを有するクラスタが行っていたディスク装置にアクセスする必要がある処理を行い、そのクラスタのキャッシュメモリのライトデータの冗長化のみを他の正常動作のクラスタに行わせる。

【選択図】 図 8

認定・付加情報

特許出願の番号	特願 2003-199581
受付番号	50301201563
書類名	特許願
担当官	第七担当上席 0096
作成日	平成15年 7月23日

<認定情報・付加情報>

【提出日】	平成15年 7月22日
-------	-------------

特願 2 0 0 3 - 1 9 9 5 8 1

出 願 人 履 歷 情 報

識別番号

[0 0 0 0 0 5 1 0 8]

1. 変更年月日

1 9 9 0 年 8 月 3 1 日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台 4 丁目 6 番地

氏 名

株式会社日立製作所